

# An Introduction to Calculus and Algebra

LLYFRGELL  
Y COLEG NORMAL  
BANGOR.

VOLUME 3

## Algebra

Open University Set Book



2

2



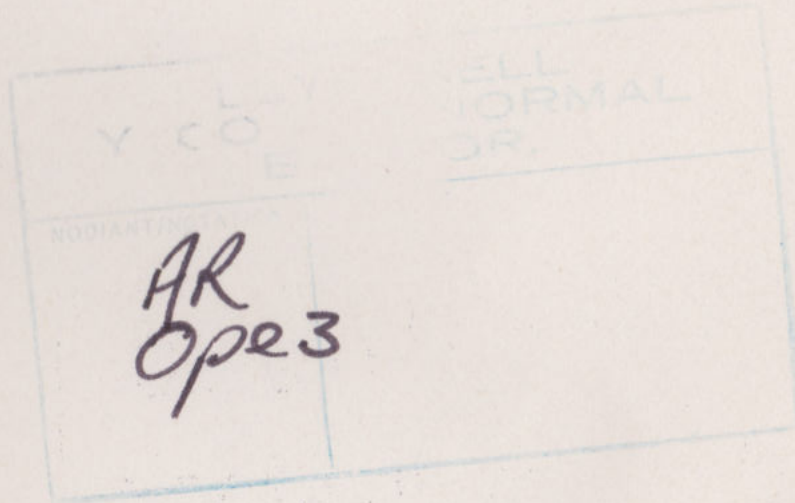
COLEG NORMAL BANGOR



3 8052 00050 011 3

# An Introduction to Calculus and Algebra

*Volume 3 Algebra*









# An Introduction to Calculus and Algebra

Volume 3

*Algebra*

The Open University Press

UWB Normal Site Library  
QA303 .O64 1971  
38052000500113

The Open University Press  
Walton Hall Bletchley  
Buckinghamshire

First published 1972  
Second impression 1972

Copyright © 1972  
The Open University

All rights reserved. No part of  
this work may be reproduced in  
any form by mimeograph or any  
other means, without permission  
in writing from the publishers.

Printed in Great Britain by  
The Staples Printing Group  
at their Woking establishment

SBN 335 00004 5



# Contents

	Page
<b>Editors' Preface</b>	ix
<b>Notation</b>	xi
<b>Chapter 1    Sets and Mappings</b>	
1.0    Introduction	1
1.1    Sets	1
1.2    Mappings	5
1.3    Functions	15
1.4    Cartesian Product	25
1.5    Additional Exercises	29
1.6    Answers to Exercises	30
<b>Chapter 2    Operations and Relations</b>	
2.0    Introduction	37
2.1    Binary Operations	37
2.2 $N$ -ary Operations	47
2.3    What is a Relation?	50
2.4    Types of Relations	56
2.5    Equivalence Relations	65
2.6    Order Relations	71
2.7    Additional Exercises	77
2.8    Answers to Exercises	79
<b>Chapter 3    Morphisms</b>	
3.0    Introduction	88
3.1    How Morphisms Arise	88
3.2    Kinds of Morphism	99
3.3    Units and Dimensions	101
3.4    Conclusion	105
3.5    Answers to Exercises	106
<b>Chapter 4    Geometric Vectors</b>	
4.0    Introduction	109
4.1    Geometric Vectors	109



4.2	Addition on the Set of Geometric Vectors	114
4.3	Scalar Multiples of Geometric Vectors	119
4.4	Linear Dependence and Independence	124
4.5	An Algebra of Number Pairs	128
4.6	“Multiplication” on the Set of Geometric Vectors	131
4.7	Applications of Geometric Vectors	137
4.8	Additional Exercises	141
4.9	Answers to Exercises	142

## Chapter 5 Vector Spaces

5.0	Introduction	151
5.1	The Algebra of Lists	151
5.2	Vector Spaces	157
5.3	Bases and Dimension of a Vector Space	164
5.4	Mapping One Vector Space to Another	166
5.5	Morphisms	175
5.6	The Kernel	179
5.7	Additional Exercises	189
5.8	Answers to Exercises	190

## Chapter 6 Matrices

6.0	Introduction	200
6.1	Linear Equations	201
6.2	Matrices	204
6.3	Combining Matrices	206
6.4	Some Special Matrices	218
6.5	Matrix Algebra and the Algebra of Numbers	221
6.6	Additional Exercises	224
6.7	Answers to Exercises	225

## Chapter 7 Linear Equations and Matrices

7.0	Introduction	230
7.1	The Nature of the Solution I	233
7.2	Solving Systems of Linear Equations	238
7.3	Systems of Linear Equations in Matrix Form	242
7.4	The Nature of the Solution II	246
7.5	The Existence Problem	249
7.6	The Uniqueness Problem	255
7.7	Summary	259
7.8	Answers to Exercises	261



## Chapter 8 Numerical Methods

8.0	Introduction	267
8.1	Elementary Matrices	267
8.2	The Inverse of a Matrix	273
8.3	Calculation of the Rank of a Matrix	278
8.4	Directive Methods	281
8.5	Iterative or Indirect Methods	295
8.6	Ill-conditioned Systems of Equations	309
8.7	Additional Exercises	315
8.8	Answers to Exercises	317

## Chapter 9 Complex Numbers

9.0	Introduction	329
9.1	A New "Square" Function	331
9.2	A New Operation on the Set of Geometric Vectors	339
9.3	The Argument	343
9.4	Real and Complex Numbers	345
9.5	Summary of Properties of Complex Numbers	351
9.6	The Algebra of Complex Numbers	353
9.7	Additional Exercises	358
9.8	Answers to Exercises	359

## Chapter 10 Complex Functions

10.0	Introduction	373
10.1	Sets of Points in the Complex Plane	373
10.2	The "Square" Function	377
10.3	Representation of Complex Functions	379
10.4	The Exponential Function	384
10.5	The Function $z \mapsto \frac{1}{z}$	393
10.6	Composition of Complex Functions	398
10.7	The Joukowski Function	402
10.8	The "Square" Function Again	403
10.9	$n$ th Roots	410
10.10	Additional Exercises	411
10.11	Answers to Exercises	412

## Chapter 11 Second Order Differential Equations

11.0	Introduction	429
11.1	Setting up a Model	430

11.2	Finding Some Solutions	436
11.3	Finding the General Solution	439
11.4	Interpretation of the Solution	447
11.5	A Mathematical Model for Resonance	451
11.6	Interpretation of the Solution	455
11.7	Additional Exercises	460
11.8	Answers to Exercises	461
<b>Index</b>		473

# NOTE

References to particular examples or exercises are made throughout by giving chapter, section, and example or exercise number; Example 4 in Chapter 1 section 1 would thus be referred to as Example 1.1.4, and so on.



## Editors' Preface

This is the third of three volumes presenting some of the essential concepts of mathematics, a few important proofs (usually in outline), together with exercises designed to reinforce the understanding of the concepts and to develop the beginnings of technical skill.

The major part of the material used here has been selected from the correspondence texts of the *Open University Foundation Course in Mathematics*. Open University courses provide a method of study, at university level, for independent learners, through an integrated teaching system which includes textual material, radio and television programmes local tutorial arrangements, and short residential courses. The correspondence text components of the Mathematics Foundation Course were produced by a Course Team (the names of the members are listed below) and were edited by Professor M. Bruckheimer and Dr. Joan Aldous.

The selection of material for these three volumes has been made with the needs of students of other subjects particularly in mind, by the Course Team preparing the second-year short course *Elementary Mathematics for Science and Technology*, and the three volumes constitute the set book around which the course is designed. (The members of this Course Team are listed below.)

In preparing these volumes, the Course Team has attempted to provide the kind of mathematics which is particularly useful for students who already have some knowledge of Science or Technology, but who, before proceeding in their own subjects, need to deepen their appreciation of the mathematical concepts underlying the techniques of calculus and algebra.

The special character of the original Foundation Course texts has been preserved as far as is possible, but the scope of the volumes is narrower than that of the course, which endeavours to give an overall picture of mathematics. For a much fuller appreciation of what mathematics is and what mathematics does, the reader is therefore referred to the original Mathematics Foundation Course correspondence texts.



## Mathematics Foundation Course Team

Professor M. Bruckheimer  
Dr. J. M. Aldous  
Mr. D. J. A. Burrows  
Mr. R. Clamp  
Mr. S. N. Collings  
Dr. A. Crilly  
Dr. D. A. Dubin  
Mr. H. G. Flegg  
Mr. E. Goldwyn  
Mr. N. W. Gowar  
Dr. A. Graham  
Mr. R. D. Harrison  
Mr. H. Hoggan

Mr. F. C. Holroyd  
Miss V. King  
Mr. R. J. Knight  
Dr. J. H. Mason  
Mr. R. Nelson  
Miss J. Nunn  
Professor R. M. Pengelly  
Professor O. Penrose  
Dr. G. A. Read  
Mr. J. Richmond  
Mr. E. Smith  
Professor R. C. Smith

### *Course Assistants*

Mr. J. E. Baker  
Mr. W. D. Crowe

### *General Course Consultant*

Mr. D. E. Mansfield

## Elementary Mathematics for Science and Technology Course Team

Professor R. M. Pengelly  
Mr. H. G. Flegg  
Dr. A. R. Meetham  
Mr. L. Aleeson  
Mr. G. Burt  
Dr. J. K. Cannell  
Mr. R. Clamp  
Dr. P. M. Clark  
Mr. S. N. Collings

Dr. A. Cooper  
Dr. A. Crilly  
Professor M. J. L. Hussey  
Mr. E. G. Law  
Mr. F. B. Lovis  
Mrs. V. Richards  
Dr. R. A. Ross  
Dr. T. B. Smith

### *Course Assistant*

Mr. R. W. Duke



# Notation

		Page
$a \in A$	$a$ is an element of the set $A$ (" $a$ belongs to $A$ ")	2
$\{x: x \text{ has a property } P\}$	The set of <i>all</i> elements $x$ which have the given property $P$	3
$\{a, b, c, d, \dots\}$	The set of elements $a, b, c, d, \dots$	2
$A \subset B$	The set $A$ is a proper subset of set $B$	4
$A \subseteq B$	The set $A$ is subset of the set $B$	4
$\emptyset$	The empty set (i.e. the set containing no elements)	4
$A = B$	The sets $A$ and $B$ have the same elements	4
$f: A \longrightarrow B$	The mapping $f$ maps the set $A$ to the set $B$	5
$f: x \longmapsto y (x \in A)$	The mapping $f$ has domain $A$ and assigns to $x \in A$ the image $y$	6
$f(x)$	The image of $x$ under the mapping $f$	7
$f(A)$	The set of images of the elements of $A$ under $f$	7
$f + g$	The sum of two functions $f + g: x \longmapsto f(x) + g(x)$	15
$g \circ f$	The function defined by $g \circ f: x \longmapsto g(f(x))$ ( $x \in \text{domain of } f$ )	18
$P \times Q$	The Cartesian product of sets $P$ and $Q$	26
$A \cap B$	The set of all elements which belong <i>both</i> to set $A$ and to set $B$ (" $A$ intersection $B$ ")	44
$A \cup B$	The set of elements which belong to set $A$ <i>or</i> to set $B$ <i>or</i> to both (" $A$ union $B$ ")	44
$A'$	The complement of the set $A$	50
$a \rho b$	$a$ is related to $b$	53
$A/\rho$	The quotient set (read as " $A$ by $\rho$ ")	68
$\left. \begin{matrix} < \\ \leq \end{matrix} \right\}$	The symbols of ordering	72
$(A, \circ)$	The set $A$ together with the binary operation $\circ$ defined on $A$	96
$\vec{a}$	An arrow	111
$\overrightarrow{AB}$	The arrow which has its tail end at $A$ , length $AB$ , and direction from $A$ to $B$	111
$\underline{AB}$	The geometric vector to which $\overrightarrow{AB}$ belongs	112
$\underline{a}$	The geometric vector to which $\vec{a}$ belongs	112
$\vec{0}$	The zero geometric vector	117
$ \underline{a} $	The length of $\underline{a}$	119
$\underline{a} \cdot \underline{b}$	The inner product of $\underline{a}$ and $\underline{b}$	132
$(a_1, a_2, \dots, a_n)$	The list of elements $a_1, a_2, \dots, a_n$	152
$D$	The differentiation operator	154
$f'$	The derived function of $f$	154
$\underline{v}_1, \underline{v}_2, \dots, \underline{v}_k$	Elements of a vector space	158
$R^n$	The Cartesian product set $\underbrace{R \times R \times \dots \times R}_{n \text{ terms}}$	168
$A$	The matrix $A$ , and also the linear transformation defined by $A$ .	205



		Page
$R_\theta$	The matrix corresponding to the mapping which rotates the plane about the origin through an angle $\theta$ anti-clockwise, i.e. $\begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}$	219
$I$	The identity matrix	221
$I_{n,n}$	The identity matrix of order $n \times n$	221
$r(A)$	The rank of the matrix $A$	252
$(A \ \underline{b})$	The augmented matrix for the system $A\underline{x} = \underline{b}$	255
$A^{-1}$	The inverse of $A$ , where $A$ is a non-singular matrix; also the inverse mapping, where $A$ is an isomorphism	259
$E_i$	An elementary matrix	268
$\underline{x}^{(r)}$	The $r$ th approximation to the solution vector in an iterative method	298
$\underline{e}^{(r)}$	The error vector: $(\underline{x}^{(r)} - \underline{X})$	298
$  \underline{a}  $	A norm of the vector $\underline{a}$ defined by $  \underline{a}   =  a_1  +  a_2  + \cdots +  a_n ,$ where $\underline{a}$ is an $n \times 1$ vector with elements $a_i$ ( $i = 1, \dots, n$ )	304
$(r, \theta)$	The polar co-ordinates of a point (or a complex number)	334
$[a, b[$	$\{x: x \in R, a \leq x < b\}$	337
$\oplus_{2\pi}$	Addition modulo $2\pi$	339
$\otimes$	The operation on the set of Cartesian co-ordinates which corresponds to multiplication on the set of polar co-ordinates	341
$p$	The many-one mapping $p: (r, \theta) \longmapsto (r \cos \theta, r \sin \theta) \quad ((r, \theta) \in R_0^+ \times R)$	343
$P$	The one-one mapping $P: (r, \theta) \longmapsto (r \cos \theta, r \sin \theta) \quad ((r, \theta) \in A)$	343
$\arg(x, y)$	The argument of $(x, y)$	343
$\text{Arg}(x, y)$	The principal value of the argument of $(x, y)$	343
$C$	The set of all complex numbers	346
$\left. \begin{matrix} x + iy \\ z \end{matrix} \right\}$	The complex number $(x, y)$	349
$\text{Re } z$	The real part of $z$	350
$\text{Im } z$	The imaginary part of $z$	350
$ z $	The modulus of $z$	355
$\bar{z}$	The conjugate of $z$	355
$\exp$	The complex exponential function: $\exp: z \longmapsto e^x(\cos y + i \sin y) \quad (z \in C),$ where $z = x + iy$	386
$D^n$	$\underbrace{D \circ D \circ \cdots \circ D}_{n \text{ terms}}$	443



# CHAPTER 1 SETS AND MAPPINGS

## 1.0 Introduction

In this first chapter our main concern is to explain and define the terms *set*, *mapping*, *function* and *Cartesian product*. (We have already discussed these concepts in Volume 1, Chapters 1 and 3, but we are repeating much of the material of those chapters here in a condensed form, since the material is basic to algebra as well as to calculus, and it therefore makes this volume somewhat more self-contained. There is no need to read this chapter if you have read Volume 1, but it may prove useful for revision and reference.)

Once the general idea of a set and a mapping is grasped, there is a need for some appropriate notation. The notation which we use here is explained and some further concepts introduced, so that precise definitions can be given. The particular kind of mapping which we call a *function* is fundamental to mathematics and we therefore devote a whole section of the chapter to functions.

The discussion of *Cartesian product*, with which we conclude the chapter, provides an opportunity to look at mapping and function from a slightly different standpoint, and involves the concept of *ordered pair* which we shall need frequently in subsequent discussions.

## 1.1 Sets

Mathematicians use the word *set* in a way which is very similar to its use in ordinary everyday speech. Roughly speaking, a set is a collection of objects. We do, however, need to be a little more precise and it is customary to define a set as **a collection of distinct well-defined objects**. This definition emphasizes two important properties of a set:

- (i) no two objects belonging to a given set are identical,
- (ii) for any object whatever, we can say whether or not it belongs to a given set.

Often, so that we can refer to them more easily, we shall denote sets by capital letters, such as  $A, B, \dots, X, Y, \dots$ . For example,  $Z$  denotes the set of all integers,  $R$  the set of all real numbers.



*Example 1*

The *Highway Code* gives the following table for stopping distances of a car travelling at various speeds.

Speed (mile/h)	20	30	40	50	60
Stopping Distance (ft)	40	75	120	175	240

We could call the set of numbers representing speeds  $A$  and the set of numbers representing stopping distances  $B$ , so that

$A$  is the set  $\{20, 30, 40, 50, 60\}$

and  $B$  is the set  $\{40, 75, 120, 175, 240\}$ .

Notice the way we write them: we list all the numbers of the set (*in any order*), separate them with commas, and enclose the whole lot with “curly brackets” (braces).

An object belonging to a set is called an **element** of the set (or sometimes a **member** of the set). We use a lower case letter to stand for an element, so we get statements like

“ $x$  is an element of  $X$ ”

or just

“ $x$  belongs to  $X$ ”.

We use statements like this so often that we find it convenient to have a symbol in place of the words

“is an element of”.

The symbol we use is  $\in$ , so that

“ $x$  is an element of  $X$ ” becomes “ $x \in X$ ”

but we still read it in the same way as before, or use the words “ $x$  belongs to  $X$ ”.

There is a standard way of defining and denoting sets which have no conventionally accepted symbol to represent them.

A typical form of words is

“the set of all  $x$  such that  $x$  has some property”



This is conventionally written in mathematical shorthand as follows:

$$\{x : x \text{ has some property}\}$$

For example:

$$\{x : x \text{ is a person with blue eyes}\}$$

which we would read as:

The set of *all*  $x$  **such that**  $x$  is a person with blue eyes

Notice that the **colon** in the notation corresponds to “**such that**” in the reading.

For sets of numbers, we have, for example:

$$\{x : x \in R \text{ and } x > 2\}$$

which we read as:

The set of all  $x$  such that  $x$  belongs to the set of all real numbers and  $x$  is greater than 2

or, more briefly, as:

The set of all real numbers greater than 2

Sometimes it is convenient to refer to a particular set by listing all its elements (as we did in Example 1), but we can, of course, do this only if the number of elements is finite. The order in which the elements are listed is immaterial since we are interested in the set as a whole, and so in Example 1 we could equally well write

$A$  as  $\{20, 50, 30, 60, 40\}$   
instead of  $\{20, 30, 40, 50, 60\}$ .



## Equality of Sets

Two sets are said to be **equal** if they contain the same elements. Thus we write

$$\{1, 2, 3\} = \{3, 1, 2\}.$$

## Subsets

Any set of elements chosen from a set is called a **subset**. For example:

- (i)  $\{367, 20, 260\}$  is a subset of  $\{367, 248, 17, 18, 20, 28, 177, 260\}$ ,
- (ii)  $\{a, b\}$  is a subset of  $\{a, b, c\}$ .

Strangely enough we say that

$$\{a, b, c\} \text{ is a subset of } \{a, b, c\}$$

but if we wish to say “a subset which is not just the original set” we say a **proper subset**.

To make the point clear:

$$\{a, b\} \text{ is a proper subset of } \{a, b, c\};$$

$$\{a, b, c\} \text{ is a subset, but not a proper subset, of } \{a, b, c\}$$

We frequently use a symbol to stand for “is a subset of”. We write

$$A \subseteq B$$

to stand for “the set  $A$  is a subset of the set  $B$ ”. For “is a proper subset of”, we write

$$A \subset B$$

to indicate that  $A$  is a subset of  $B$  and that  $A$  is not equal to  $B$ .

Sometimes we want to refer to a rather special set, the set **having no elements** which we call the **empty set** (or **null set**). We denote this set by  $\emptyset$ . The **empty set is a subset of every set**.

### Exercise 1

In each of the following questions, indicate which (if any) statements are correct.

- (i) If  $A = \{367, 20, 260\}$  and  $B = \{367, 248, 17, 18, 20, 28, 177, 260\}$ , then

(a)  $A \subset B$

(b)  $B \subseteq A$

- (c)  $A = B$
  - (d)  $B \subset A$
  - (e)  $A \subseteq B$
  - (f)  $A$  is a proper subset of  $B$
  - (g)  $A$  is a subset of  $B$
- (ii) If  $A = \{\text{Jim, Mary}\}$  and  $B = \{\text{Blue, Green}\}$ , then
- (a)  $A = B$
  - (b)  $A \subset B$
  - (c)  $B \subset A$

## 1.2 Mappings

In Example 1.1.1 we had two related sets of numbers which we called  $A$  and  $B$ . We can say that

To each number in  $A$  a number in  $B$  is assigned

or

The set  $A$  is mapped to the set  $B$

or, in symbols,

$A \longrightarrow B$

which we read as “ $A$  maps to  $B$ ”.

We can now make a first attempt at a definition of the term “mapping”:

The essential feature of a **MAPPING** is that we have two sets and a method of assigning to *each* element of one set one or more elements of the other.

### Example 1

We can map the set of all towns in the British Isles to the set of points on a piece of paper, by drawing a geographical map of the British Isles on the paper.

### Example 2

We can map the set of all people to the set of all integers, using the rule that to each person is assigned his height in centimetres measured to the nearest centimetre.



(Notice that there are numbers in the second set in this example which are not assigned to any person in the first set. We don't know of anybody 2 cm. high.)

It is convenient to have a shorthand notation for statements like "A speed of 20 mile/h maps to a stopping distance of 40 ft".

We already have the notation

$$A \longrightarrow B$$

which stands for "set  $A$  maps to set  $B$ ".

What we need is a notation which says that a *particular* element of  $A$  has assigned to it a *particular* element or set of elements of  $B$ .

If  $a$  belongs to  $A$  (i.e.  $a \in A$ ) and  $b$  belongs to  $B$  (i.e.  $b \in B$ ), the statement " $b$  is assigned to  $a$ " is abbreviated to

$$a \longmapsto b$$

Instead of the precise " $b$  is assigned to  $a$ " we also often say " $a$  maps to  $b$ ", where the context makes clear which we mean.

Notice that we have just added a small bar to the arrow to indicate that this is a precise assignment of elements rather than just a mapping of one set to another, where there may be elements in the second set which are not assigned to elements in the first.

## Images

If  $a \in A, b \in B$

and  $a \longmapsto b$ ,

we say that  $b$  is the IMAGE of  $a$ .

The image of an element may be a set of elements. For example, if we map natural numbers to their factors, then we shall have, for example,

$$6 \longmapsto \{1, 2, 3, 6\}.$$

We could consider an image to be a set even if it consisted of only one element, but we do not make the distinction in this context between an element and the set comprising that single element.



### Naming a Mapping

If we wish to refer to a mapping, it is not always convenient to have to describe it in full every time, so we often use a symbol to represent it. For example, the mapping of the set of all people to heights in centimetres (Example 2) could be represented by the letter  $h$ . We then write

$$h: \text{Set of people} \longrightarrow \text{Set of heights in centimetres}$$

involving whole sets, and

$$h: \text{John Brown} \longmapsto 183 \text{ cm}$$

involving elements of sets. Using the letter  $h$  we also write  $h$  (John Brown) to mean the image of John Brown under the mapping  $h$ . Therefore, we have

$$h(\text{John Brown}) = 183 \text{ cm.}$$

### Summary of Notation

$a \in A$  means  $a$  is an **element** of set  $A$ .  
 $m: A \longrightarrow B$  means The **mapping**  $m$  maps the set  $A$  to the set  $B$ .  
 $m: a \longmapsto b$  means The **mapping**  $m$  maps the element  $a$  to the **element**  $b$ ,  
 OR  
 $b$  is the **image** of  $a$  under (the mapping)  $m$ .

If the need arises we extend the notation still further and write,

$$h: A \longmapsto B \text{ or } h(A) = B$$

when every element of the set  $B$  corresponds to at least one element of the set  $A$ .

### Example 3

The following tables are all typical examples of results which are obtained from experiment or observation.

Table I

$s$	Speed (mile/h)	20	30	40	50	60
	Distance (ft)	40	75	120	175	240



Table II

$f$	Year	1938	1939	1940	1941	1942	1943	1944	1945	1946
	Sale	367	248	17	18	20	20	28	177	260

Table III

$t$	Distance	2	4	6	8	10
	Temperature	25	42	50	51	44

The image of 20 under  $s$  is 40 and so we write

$$s: 20 \longmapsto 40$$

or

$$s(20) = 40$$

The image of 1943 under  $f$  is 20 and so we write

$$f: 1943 \longmapsto 20$$

or

$$f(1943) = 20$$

The image of  $\{2, 4, 6\}$  under  $t$  is  $\{25, 42, 50\}$  and so we write

$$t: \{2, 4, 6\} \longmapsto \{25, 42, 50\}$$

or

$$t(\{2, 4, 6\}) = \{25, 42, 50\}$$

If  $A = \{2, 4, 6, 8, 10\}$  and  $B = \{25, 42, 50, 51, 44\}$  we can write

$$t: A \longmapsto B \quad \text{and} \quad t(A) = B$$

But if  $C = \{25, 42, 50, 51, 44, 99\}$ , we *cannot* write  $t: A \longmapsto C$  or  $t(A) = C$  because 99 has no corresponding element in  $A$ , but we *can* write

$$t: A \longrightarrow C \quad \text{or} \quad t(A) \subseteq C$$

Exercise 1

Complete each of the following (i.e. replace the question mark appropriately, where  $s, f, t$  are the mappings in Example 3).

- (i) The image of 30 under  $s$  is ?

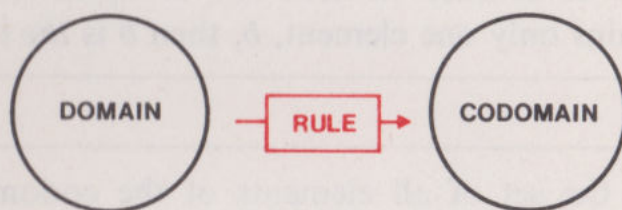
- (ii)  $t:8 \mapsto ?$
- (iii)  $f: ? \mapsto 28$
- (iv)  $f(1942) = ?$
- (v)  $s(60) = ?$
- (vi)  $t(?) = 44$
- (vii)  $s:\{30, 40, 50\} \mapsto ?$

### Domain and Codomain

There are several points still to tidy up if we wish to be more precise about the meaning of words such as MAPPING.

If we have a mapping from a set  $A$  to a set  $B$ , then we call  $A$  the **DOMAIN** of the mapping, and  $B$  the **CODOMAIN** of the mapping.

The mapping involves some method, or rule, whereby to **each** element of the domain an image is assigned, and the codomain contains all the images.



In Example 1 the domain is:

the set of all towns in the British Isles.

The codomain is:

the set of points on a piece of paper.

In our first tentative definition of a mapping (page 5) we required that to each element of the domain we must be able to assign an element (or elements) of the codomain. In other words each element of the domain must have an image in the codomain. On the other hand, there is no



reason why the codomain should not include elements which are not images, for there is no question of applying a formula, or a rule, to elements of the codomain. Thus, all we require of the codomain is that it should contain the set of all images of elements in the domain.

We are now in a position to list some definitions. *Notice that in the first definition we define a mapping to consist of three things (the domain, the rule, and a codomain).*

A **MAPPING** consists of a set  $A$ , a set  $B$  and a rule by which an element (or set of elements) of  $B$  is assigned to *each* element of  $A$ .

The set  $A$  is the **DOMAIN** of the mapping.

The set  $B$  is the **CODOMAIN** of the mapping.

If an element  $a$  of the domain has assigned to it an element  $b$  or a set of elements  $T$  of the codomain, then  $b$  or  $T$  is the **IMAGE** of  $a$ . Each element of  $T$  is called *an* image of  $a$ . If  $T$  contains only one element,  $b$ , then  $b$  is *the* image of  $a$ .

If  $T$  is the set of all elements of the codomain which are images of elements of the subset  $S$  of the domain, then  $T$  is the **IMAGE** of  $S$ .

The mappings for which each element in the domain has a unique element as its image are particularly important, and so we give them a special name.

A **FUNCTION** is a mapping for which each element in the domain has only one element as its image.



*Note*

It is becoming increasingly common usage to use the terms “mapping” and “function” synonymously. What we have called a mapping is sometimes referred to as a “correspondence”. We have kept to our definition of mapping because we feel that the word mapping carries with it more of a sense of movement from one set to another than the word correspondence. As with any mathematical term, you must be sure, if you consult a text book of the way in which the author is using the term.

*Exercise 2*

In each case state whether the statement is true or false:

- (i) A mapping is always a function.
- (ii) The domain of a mapping is the set of all images under the mapping.
- (iii) If  $m$  is a function with domain  $A$  and codomain  $B$ , then  $m(a)$  must be an element of  $B$ . ( $a \in A$ )
- (iv) The set of all images under a mapping is called the codomain of the mapping.
- (v) If  $A = \{\alpha, \beta, \gamma\}$  and  $B = \{1, 2, 3\}$  then the list:

$$m: \alpha \longmapsto \{1, 2\}$$

$$m: \beta \longmapsto 1$$

$$m: \gamma \longmapsto 1$$

defines a mapping from  $A$  to  $B$ .

- (vi) The list in (v) defines a function.
- (vii) If  $A = \{\alpha, \beta, \gamma\}$  and  $B = \{1, 2, 3\}$  then the list:

$$m: \alpha \longmapsto \{1, 2\}$$

$$m: \beta \longmapsto 3$$

defines a mapping from  $A$  to  $B$ .

When the domain and codomain of a mapping are sets of numbers, we can often abbreviate the rule which tells us how to find the image of the domain by using common algebraic notation.



*Example 4*

Let  $f$  be the mapping with domain and codomain the set of all real numbers, and the rule:

for any real number (i) square it  
 (ii) multiply the result by 6  
 (iii) subtract from this result twice the original number  
 and then (iv) add 1

We usually abbreviate this and refer to the mapping

$$f: x \longmapsto 6x^2 - 2x + 1 \quad (x \in R)$$

Any letter could have been used in place of  $x$  as an arbitrary element of the domain; however when the domain is  $R$  it is common practice to use  $x$ . Any letter used in this way, to make a statement about every element of a set, is called a **VARIABLE**.

The statement  $x \in R$ , in the formula above, contains rather a lot of information. It tells us that

(i) the domain is  $R$

and

(ii) the letter  $x$  is a variable, which can take any value in the domain.

Notice that the mapping is a *function* because the value of  $x$  defines the value of  $6x^2 - 2x + 1$  *uniquely*. Any letter could have been used to define this function. Thus

$$f: t \longmapsto 6t^2 - 2t + 1 \quad (t \in R)$$

and  $f: a \longmapsto 6a^2 - 2a + 1 \quad (a \in R)$

define exactly the same function. On the other hand, the statements

$$f: x \longmapsto 6x^2 - 2x + 1$$

and

$$f: 2 \longmapsto 6 \times 2^2 - 2 \times 2 + 1 = 21$$

do not define a mapping, for there is no mention of the domain. They



are simply statements about what happens to *particular* elements under the mapping, in the first case the element  $x$ , in the second the element 2. Our definition of a mapping states that we must (among other things) specify the codomain, and yet we seem to be ignoring it. Is it true that

$$f: x \longmapsto x^2 + 4x - 1 \quad (x \in R)$$

defines a mapping?

Strictly speaking the answer is NO, but if we assume that the codomain is also  $R$  then the definition is complete. Since, in general, given the domain and the rule we can work out the set of images (and the codomain can be any set containing the set of images), we shall not include the specification of the codomain in the definition of the mapping unless we have a particular interest in it.

Strictly speaking mappings are **equal** only if they have the same DOMAIN, CODOMAIN and RULE. However, we shall be content with having the same domain and rule. For example, in the mapping above

$$f: x \longmapsto x^2 + 4x - 1 \quad (x \in R)$$

we would not distinguish between the two cases where the codomain is  $R$  and where it is the set of real numbers greater than or equal to  $-5$ , which is the set of images. If  $f$  and  $g$  are two mappings with the same domain and rule, we shall write

$$f = g$$

### Exercise 3

For each of the functions  $f$ ,  $g$  and  $h$  defined below, state

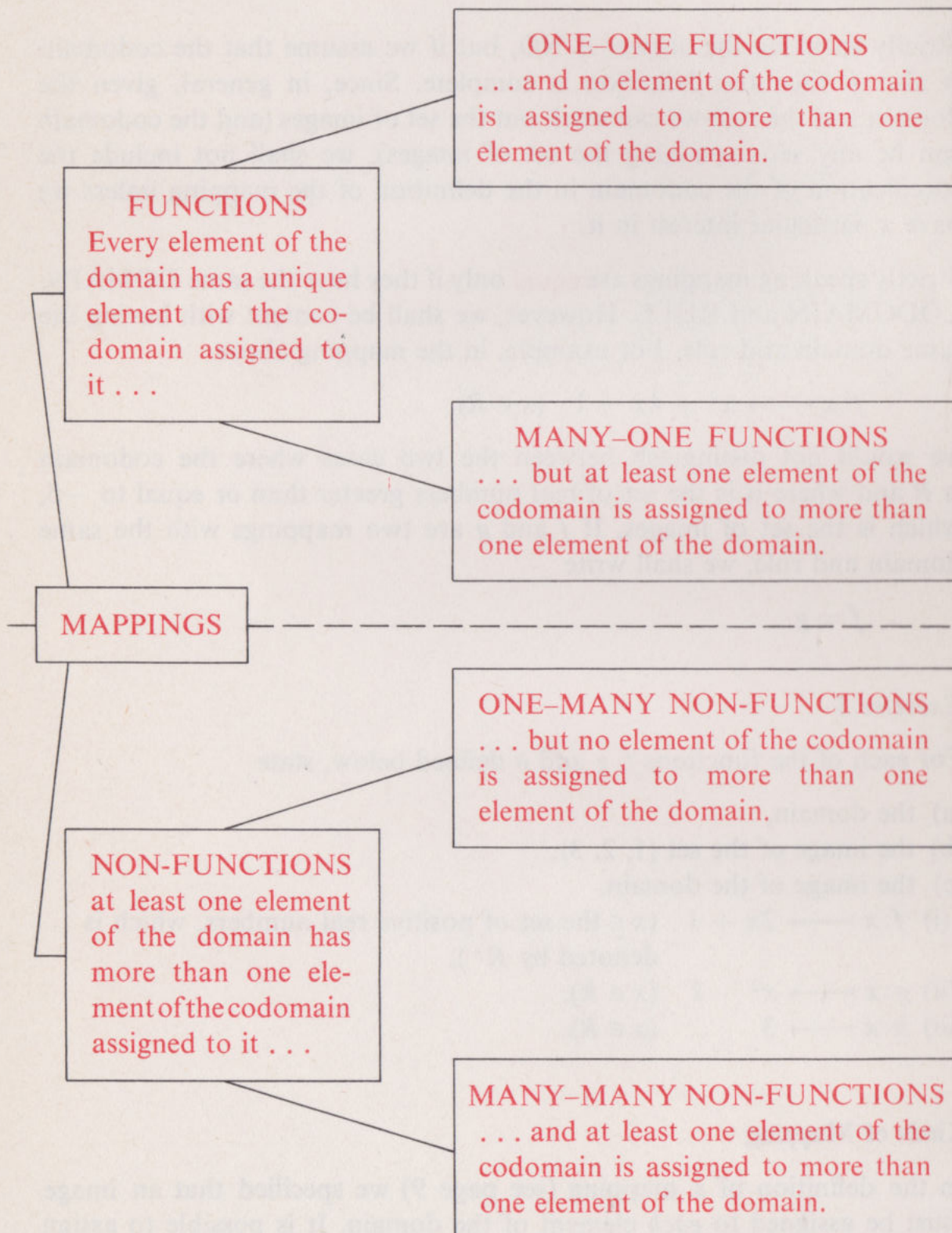
- (a) the domain,
- (b) the image of the set  $\{1, 2, 3\}$ ,
- (c) the image of the domain.
  - (i)  $f: x \longmapsto 2x + 1 \quad (x \in \text{the set of positive real numbers, which is denoted by } R^+)$ ,
  - (ii)  $g: x \longmapsto x^2 - 2 \quad (x \in R)$ ,
  - (iii)  $h: x \longmapsto 3 \quad (x \in R)$ .

### Kinds of Mapping

In the definition of a mapping (see page 9) we specified that an image must be assigned to *each* element of the domain. It is possible to assign



images in various ways. As we have already seen, if each element in the domain has only one element of the codomain as its image, then the mapping is called a *function*. Again there are some functions where no element of the codomain is assigned to more than one element of the domain, and others where several elements of the domain have the same image. The various possibilities are summarized in the following diagram:



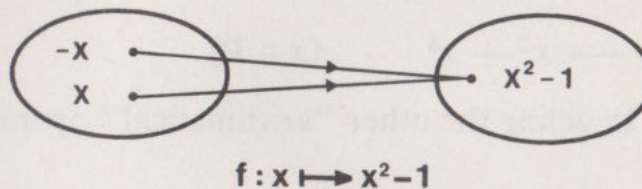


*Example 5*

The mapping defined by

$$f: x \mapsto x^2 - 1 \quad (x \in R)$$

is many-one (a function).  $x^2 - 1$  is defined as just one number when  $x$  is given a value. On the other hand, an image may correspond to more than one number; for example,  $f(3) = 8$  and  $f(-3) = 8$ .

*Exercise 4*

Classify the following mappings as

one-one or many-one or one-many or many-many

- (i)  $x \mapsto 3x^2 + 2$   $(x \in R)$
- (ii)  $x \mapsto x^3 + 2$   $(x \in R)$
- (iii)  $x \mapsto \sin x$   $(x \in R)$
- (iv)  $x \mapsto \{x, -x\}$   $(x \in R^+)$
- (v)  $x \mapsto \{\sqrt{1 + 3x^2}, -\sqrt{1 + 3x^2}\}$   $(x \in R)$

**1.3 Functions**

We have seen in section 1.2 that a function is distinguished from other mappings by every element of the domain having a unique element of the codomain assigned to it. Let us now see how functions can be combined together.

**The “Arithmetic” of Functions**

Suppose two functions  $f$  and  $g$  have  $R$  as domain and codomain. Then we can define

the **SUM** of  $f$  and  $g$

which we write as  $f + g$ , by

$$f + g: x \mapsto f(x) + g(x) \quad (x \in R)$$



*Example 1*

Let

$$f: x \longmapsto x^2 \quad (x \in R)$$

and

$$g: x \longmapsto x^6 \quad (x \in R)$$

Then

$$f + g: x \longmapsto x^2 + x^6 \quad (x \in R)$$

It is quite natural to define the other “arithmetical” operations as follows:

**DIFFERENCE**

$$f - g: x \longmapsto f(x) - g(x) \quad (x \in R)$$

**PRODUCT**

$$f \times g: x \longmapsto f(x) \times g(x) \quad (x \in R)$$

**QUOTIENT**

$$f \div g: x \longmapsto \frac{f(x)}{g(x)}$$

The specification of the domain of the quotient is not straightforward. This is because of the difficulty which occurs when  $g(x) = 0$ . In this case the image of  $x$  is undefined, and we must remove such elements from the domain. So the domain of  $f \div g$  is  $R$  with these elements omitted.

*Exercise 1*

If the functions  $f$  and  $g$  are defined by

$$f: x \longmapsto 6x^2 \quad (x \in [-1, 1])^*$$

and

$$g: x \longmapsto 6x \quad (x \in [-1, 1])$$

fill in the formula and the appropriate domain for

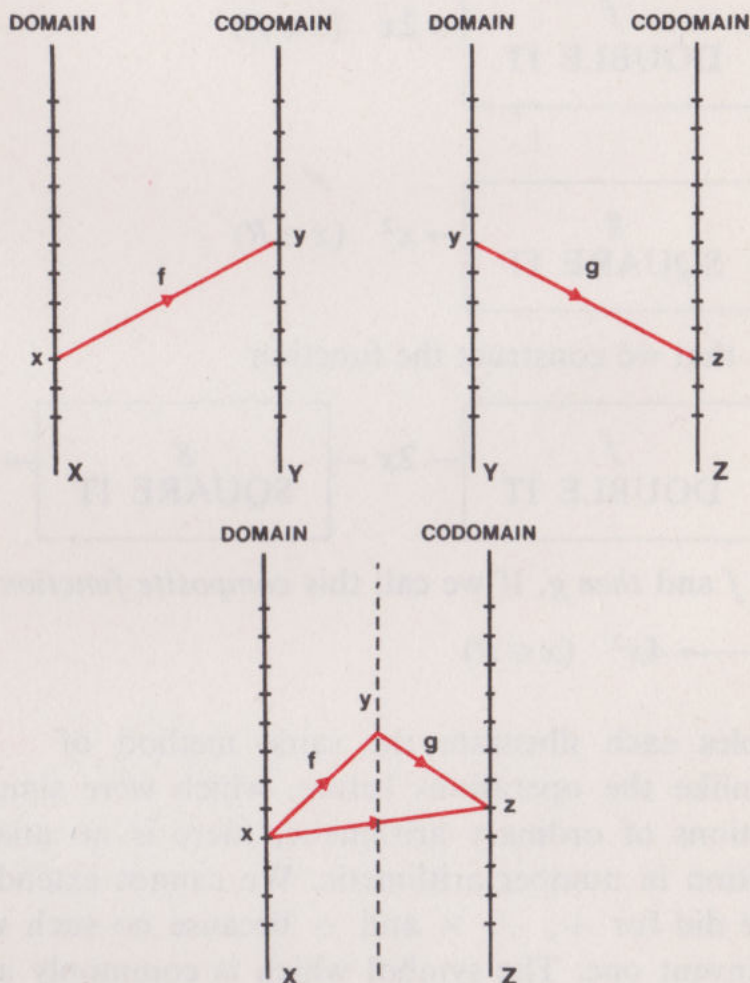
- (i)  $g + f$
- (ii)  $g \div f$
- (iii)  $f \div g$
- (iv)  $f \times g$

\* By  $[a, b]$  we mean the set of real numbers  $x$  such that  $a \leq x \leq b$ .

## Composition of Functions

There is another way of combining functions which is fundamentally different from the “arithmetical” combinations. The emphasis of this combination is on the mapping from one set to another, rather than being a simple generalization of the operations of ordinary arithmetic.

### Example 2



Notice that in the above example

$$f: X \longrightarrow Y \quad \text{and} \quad g: Y \longrightarrow Z$$

The mapping in the bottom figure is obtained by using first  $f$  and then  $g$ . If we call it  $h$ , then

$$h: X \longrightarrow Z$$



## Example 3

Suppose that we have the functions

$$f = \boxed{\text{DOUBLE IT}} \text{ with domain } R$$

$$g = \boxed{\text{SQUARE IT}} \text{ with domain } R$$

then

$$x \mapsto \boxed{\begin{array}{c} f \\ \text{DOUBLE IT} \end{array}} \rightarrow 2x \quad (x \in R)$$

and

$$x \mapsto \boxed{\begin{array}{c} g \\ \text{SQUARE IT} \end{array}} \rightarrow x^2 \quad (x \in R)$$

Suppose now that we construct the function

$$x \mapsto \boxed{\begin{array}{c} f \\ \text{DOUBLE IT} \end{array}} \rightarrow 2x \rightarrow \boxed{\begin{array}{c} g \\ \text{SQUARE IT} \end{array}} \rightarrow 4x^2$$

by using *first*  $f$  and *then*  $g$ . If we call this *composite function*  $h$ , then

$$h: x \mapsto 4x^2 \quad (x \in R)$$

These examples each illustrate the same method of **composition of functions**. Unlike the operations before, which were simply extensions of the operations of ordinary arithmetic, there is no analogue of this new composition in number arithmetic. We cannot extend the use of a symbol as we did for  $+$ ,  $-$ ,  $\times$  and  $\div$  because no such symbol exists; so we must invent one. The symbol which is commonly adopted is the small circle  $\circ$ .

Thus  $g \circ f$  (pronounced “gee oh eff”) stands for the function obtained by performing  $f$  *first*, and *then*  $g$ .

If we can define a function  $h$  by the rule:

$$h(x) = g(f(x)) \quad (x \in \text{domain of } f)$$

then we denote this function by

$$h = g \circ f$$

(The only reason for introducing  $h$  into this definition is simply that we thought that

$$(g \circ f)(x) = g(f(x))$$

for all  $x$  in the domain of  $f$

is not quite as clear. Sometimes  $g \circ f$  is referred to as a *function of a function*.)

Diagrammatically we have the extended rule

$$x \longmapsto f(x) \longmapsto g(f(x))$$

It is very important to notice that  $g \circ f$  means that we use  $f$  first and then  $g$

#### Example 4

If functions  $f$  and  $g$  are defined by

$$f: x \longmapsto 2x + 3 \quad (x \in R)$$

and

$$g: x \longmapsto x^2 - 1 \quad (x \in R)$$

we can calculate  $g(f(x))$  by replacing  $x$  by  $f(x)$  in the expression for  $g(x)$ :

$$g(x) = x^2 - 1$$

and so

$$g(f(x)) = [f(x)]^2 - 1$$

But

$$f(x) = 2x + 3$$

and so

$$\begin{aligned} g(f(x)) &= (2x + 3)^2 - 1 \\ &= 4x^2 + 12x + 8 \end{aligned}$$

Thus  $g \circ f$  is the function defined by

$$g \circ f: x \longmapsto 4x^2 + 12x + 8 \quad (x \in R)$$



*Exercise 2*

(i) If  $f$  and  $g$  are functions defined by

$$f: x \longmapsto x - 1 \quad (x \in \mathbb{R})$$

and

$$g: x \longmapsto x^2 \quad (x \in \mathbb{R})$$

complete the following:

(a)  $f \circ g: x \longmapsto ? \quad (x \in \mathbb{R})$

(b)  $g \circ f: x \longmapsto ? \quad (x \in \mathbb{R})$

(ii) If  $f$  is the mapping which translates English to French and  $g$  is the mapping which translates French to German is it  $g \circ f$  or  $f \circ g$  which translates English to German?

*Exercise 3*

Given any two functions  $f$  and  $g$ ,

(i) Can we always form  $g \circ f$ ?

(ii) If we can form  $g \circ f$ , can we necessarily form  $f \circ g$ ?

Explain your answers.

**Inverse Functions**

If one is trying to locate a book in a library, then the recommended procedure is to look it up in the catalogue and find its classification number: the classification system provides a mapping,  $c$  say, such that

$$c: \text{Books} \longrightarrow \text{Classification Numbers}$$

It is highly desirable that  $c$  is a function, rather than just a mapping. Can you say why?

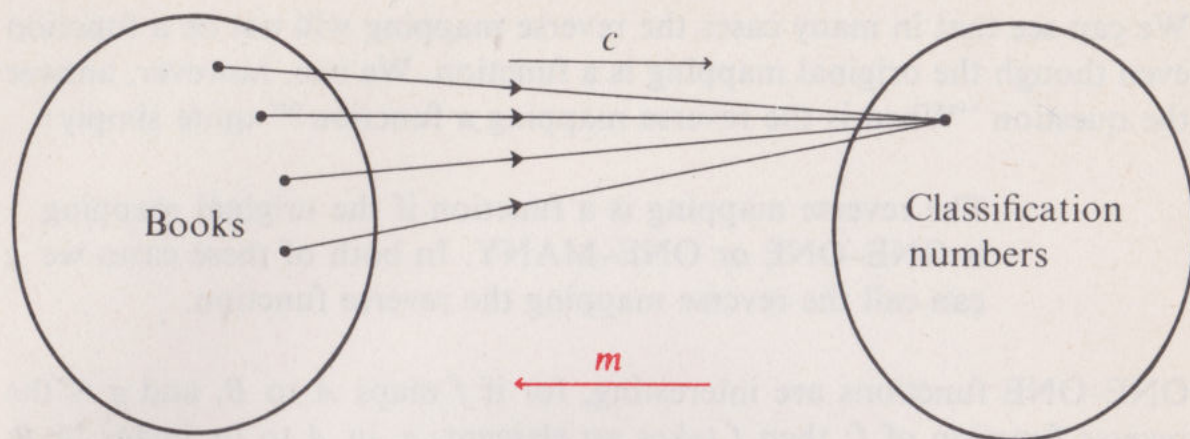
On the other hand, if you want to find a book on a particular subject, and you know the classification number of the subject, then you look in a different catalogue which represents the mapping,  $m$  say, such that

$$m: \text{Classification Numbers} \longrightarrow \text{Books}$$

If  $m$  is a function rather than a mapping, then you can be sure that the library is not much good. Can you say why?

We say that  $m$  is the **REVERSE MAPPING** to  $c$ .





Probably the most convenient way to define a reverse mapping is in terms of the list of pairs which a mapping defines. A mapping  $f$  from  $A$  to  $B$  defines a set of all the pairs  $(x, y)$  such that  $x \in A$  and  $y$  is  $f(x)$  or, if  $f(x)$  is a set of elements, belongs to  $f(x)$ . This set of pairs is called the *graph* of  $f$ .

If  $f$  maps  $A$  to  $B$  and

$$S = \{(x, y) : x \in A \text{ and } y = f(x) \text{ or } y \in f(x)\}$$

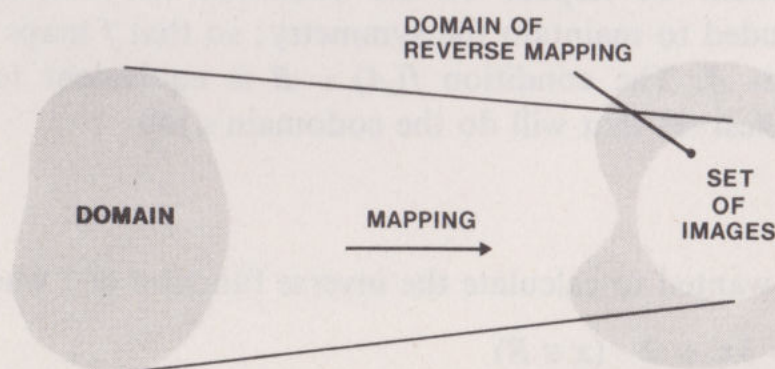
then the mapping of a subset of  $B$  to  $A$ , whose graph is

$$\{(y, x) : (x, y) \in S\}$$

is called the reverse mapping of  $f$ .

This definition is just a precise way of saying that we reverse the order of all pairs in the list which defines the mapping in order to get the reverse mapping. But one point is not entirely clear. What is the domain? We have said it is a subset of  $B$ . We now specify this subset precisely.

If  $g$  is the reverse mapping of  $f$ , and  $f$  has domain  $A$ , then the domain of  $g$  is  $f(A)$

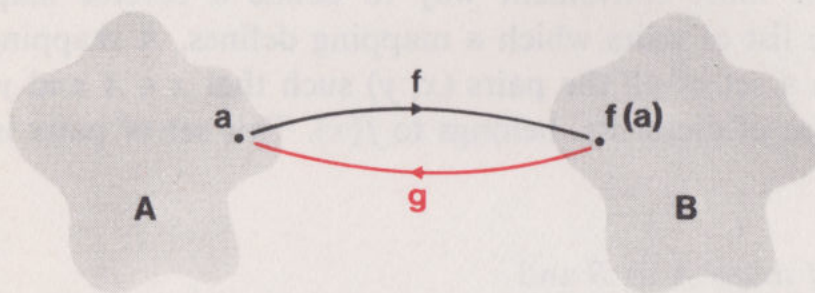




We can see that in many cases the reverse mapping will not be a function even though the original mapping is a function. We can, however, answer the question “When is the reverse mapping a function?” quite simply:

The reverse mapping is a function if the original mapping is ONE-ONE or ONE-MANY. In both of these cases we can call the reverse mapping the reverse function.

ONE-ONE functions are interesting, for if  $f$  maps  $A$  to  $B$ , and  $g$  is the reverse function of  $f$ : then  $f$  takes an element,  $a$ , in  $A$  to its image in  $B$ , and  $g$  brings this image back to  $a$  (and to  $a$  only).



In other words

$$g(f(a)) = a \quad \text{for all } a \in A$$

Notice that this statement would also be true if  $f$  were a one-many mapping. But there is a difference; for one-one mappings we can *also* state

$$f(g(b)) = b \quad \text{for all } b \in f(A)$$

We now adopt the following definition:

If  $f$  is a **ONE-ONE** function from  $A$  to  $B$ , and if  $f(A) = B$  (i.e. the codomain of  $f$  is equal to the set of all images), then the **function**  $g$  from  $B$  to  $A$  where  $g(f(a)) = a$  ( $a \in A$ ) is called the **INVERSE FUNCTION** to  $f$ .

The condition which we impose on the codomain of  $f$ , namely that  $f(A) = B$ , is included to maintain the symmetry; so that  $f$  maps  $A$  to  $B$ , and  $g$  maps  $B$  to  $A$ . The condition  $f(A) = B$  is equivalent to saying that  $B$  is the smallest set that will do the codomain's job.

### Example 5

Suppose that we wanted to calculate the inverse function of  $f$  where

$$f: x \longmapsto 3x + 2 \quad (x \in R)$$



If we put

$$y = f(x)$$

then

$$y = 3x + 2$$

and we can calculate  $y$  if we are given  $x$ .

The inverse function will enable us to calculate  $x$  if we are given  $y$ , and we can find this inverse  $g$  by rearranging the equation  $y = 3x + 2$  to give an equation of the form

$$x = \text{something involving } y \text{ (and not } x) = g(y)$$

If we do this we get

$$x = \frac{y - 2}{3}$$

and so

$$g(y) = \frac{y - 2}{3}$$

and therefore  $g$  is the mapping

$$g: y \longmapsto \frac{y - 2}{3} \quad (y \in R)$$

We could of course rewrite this in the equivalent form

$$g: x \longmapsto \frac{x - 2}{3} \quad (x \in R)$$

#### Exercise 4

Determine the inverse function of

$$f: x \longmapsto 4 - \frac{3}{x} \quad (x \in R^+)$$

(Do not forget the domain of the inverse function.)

#### Exercise 5

If  $g$  is the inverse function of the one-one function  $f$ , is it true that

- (i)  $f$  is the inverse function of  $g$ ?
- (ii)  $g \circ f = f \circ g$ ?



(iii)  $g(x) = \frac{1}{f(x)}$ ?

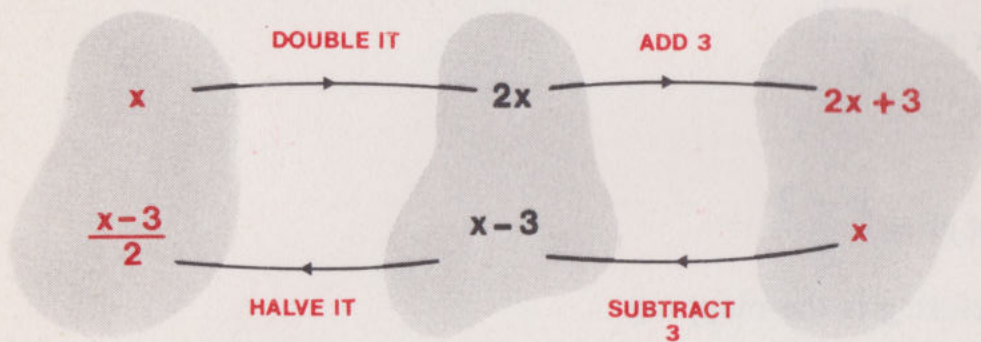
It is sometimes useful, when finding inverses of relatively simple functions, to decompose a function into more elementary ones.

### Example 6

The function

$$f: x \mapsto 2x + 3 \quad (x \in \mathbb{R})$$

has two components—“double it” and “add 3”. If we want to invert this function we must unravel the calculation; “subtract 3” and “halve it”.

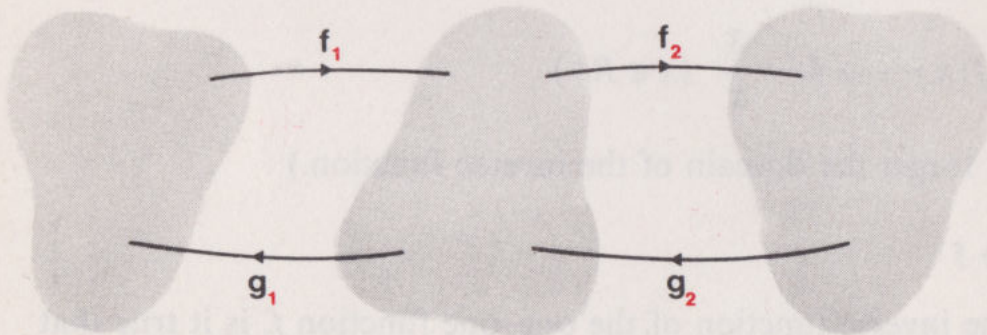


The inverse function is

$$g: x \mapsto \frac{x - 3}{2} \quad (x \in \mathbb{R})$$

In general, if  $f_1$  and  $f_2$  are one-one functions and have inverses  $g_1$  and  $g_2$  then

the inverse of  $f_2 \circ f_1$  is  $g_1 \circ g_2$



Note the order in which the inverses are combined—when inverting we have to invert the last step first.

*Exercise 6*

The one-one function  $f$  where

$$f: x \mapsto 3x^2 + 2 \quad (x \in \mathbb{R}^+)$$

maps an element  $x$  to an element  $y$ , where

$$y = 3x^2 + 2$$

The inverse function,  $g$ , will map  $y$  back to  $x$ . By changing the subject of the formula, i.e. expressing  $x$  in terms of  $y$ , find a formula for  $g$ .

## 1.4 Cartesian Product

One of the ways in which we can specify a mapping is by listing all the elements of the domain together with their images (provided that the number of elements of the domain and the image set is finite). We could, for example, represent the mapping of Example 1.1.1 by the set of pairs of numbers

$$\{(20, 40), (30, 75), (40, 120), (50, 175), (60, 240)\}.$$

Each pair consists of an element of the domain followed by its image in the codomain. Notice that the order in which the elements of any pair are written is important. Thus  $(20, 40)$  would not mean the same as  $(40, 20)$ .

When the order in which the elements of a pair are written is important, we call the pair an **ordered pair**.

*Example 1*

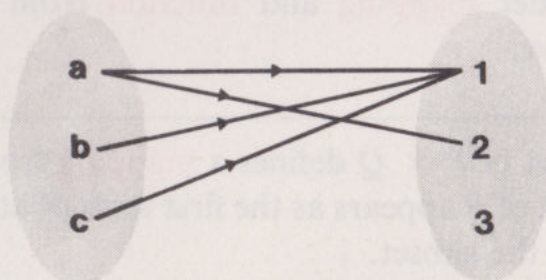
Consider the mapping  $m$  from

$$A = \{a, b, c\}$$

to

$$B = \{1, 2, 3\}$$

illustrated by





We can just as easily represent the mapping by the set of ordered pairs,

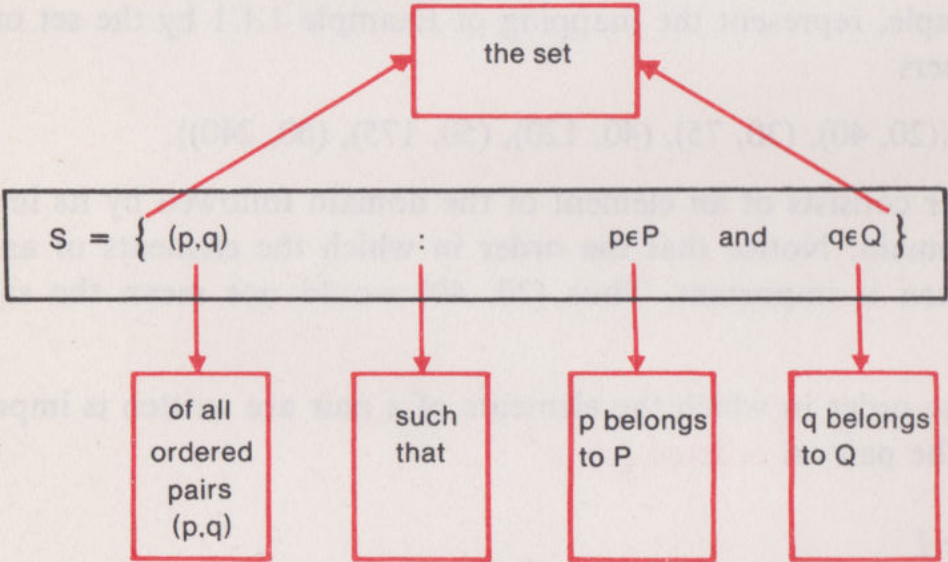
$$\{(a, 1), (a, 2), (b, 1), (c, 1)\}$$

Any mapping from  $A$  to  $B$  corresponds to a set of ordered pairs, and such a set must be a subset of all possible ordered pairs,

$$\{(a, 1), (a, 2), (a, 3), (b, 1), (b, 2), (b, 3), (c, 1), (c, 2), (c, 3)\}$$

However, our definition of mapping is such that not every subset of the set of all possible ordered pairs will define a mapping.

In general, any mapping from a set  $P$  to a set  $Q$  will correspond to a set of ordered pairs. The first element of each pair will belong to  $P$ , and the second element to  $Q$ . The set of ordered pairs representing the mapping will be a subset of the set of *all possible pairs*  $(p, q)$  where  $p \in P$  and  $q \in Q$ . If we call this *set of all possible pairs*  $S$ , we have



The set  $S$  is called the **Cartesian product** of  $P$  and  $Q$  and we denote it by

$$P \times Q$$

(We read this as “ $P$  cross  $Q$ ”.)

We can now consider **mapping** and **function** from a slightly different standpoint as follows:

A subset of  $P \times Q$  defines a **mapping** from  $P$  to  $Q$  if every element of  $P$  appears as the first term of at least one ordered pair of the subset.



A subset of  $P \times Q$  defines a **function** from  $P$  to  $Q$  if each element of  $P$  appears as the first term of an ordered pair of the subset once and once only.

(Note that the codomain  $Q$  is not necessarily defined by the subset of  $P \times Q$  in either case, because there is not a requirement that every element of the codomain should appear as the second term of at least one ordered pair of the subset. Each element of the domain and its image is specified, however, and so the set of images of the elements of  $P$  is defined, and it is this set which is of importance.)

### Example 2

Probably the most frequently occurring collection of functions is the set of functions whose domain and codomain are subsets of the set of real numbers,  $R$ . Familiar examples are:

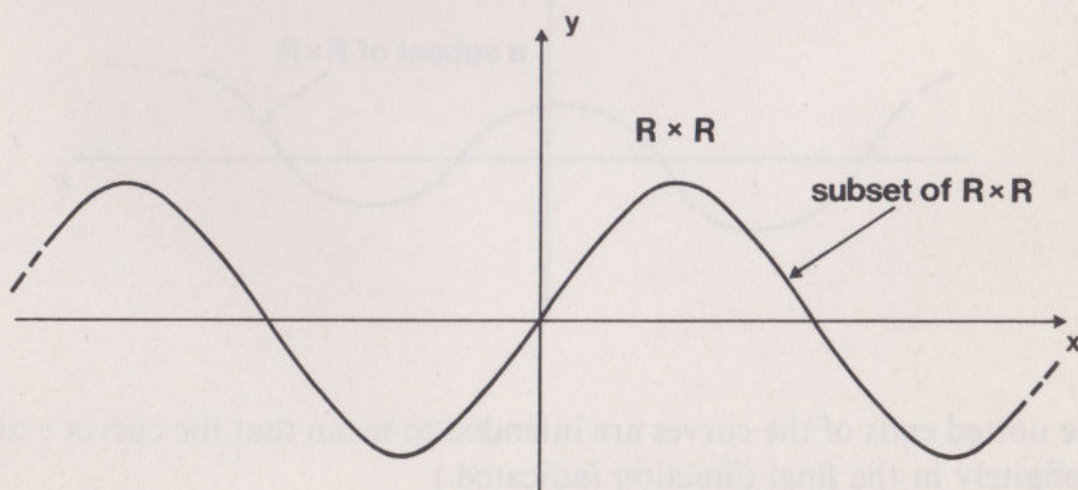
$$x \longmapsto a_1x + a_0 \quad (x \in R) \quad \text{where } a_0, a_1 \in R$$

$$x \longmapsto \sin x \quad (x \in R)$$

$$x \longmapsto \log x \quad (x \in R^+)$$

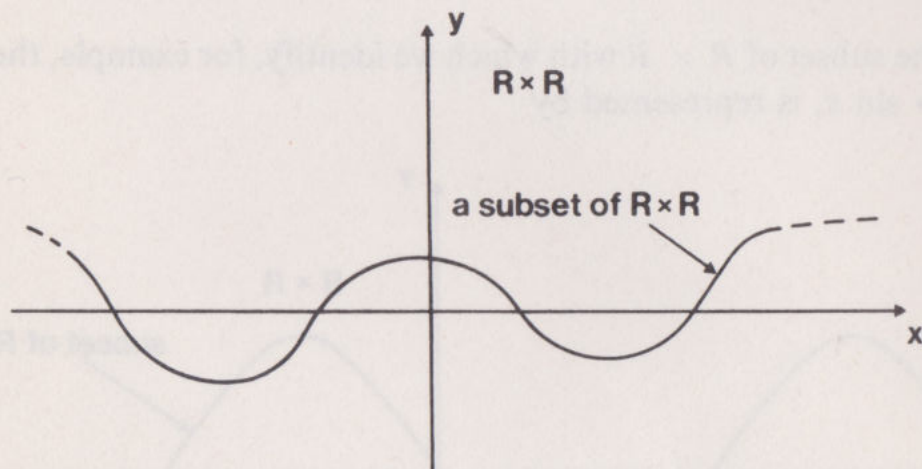
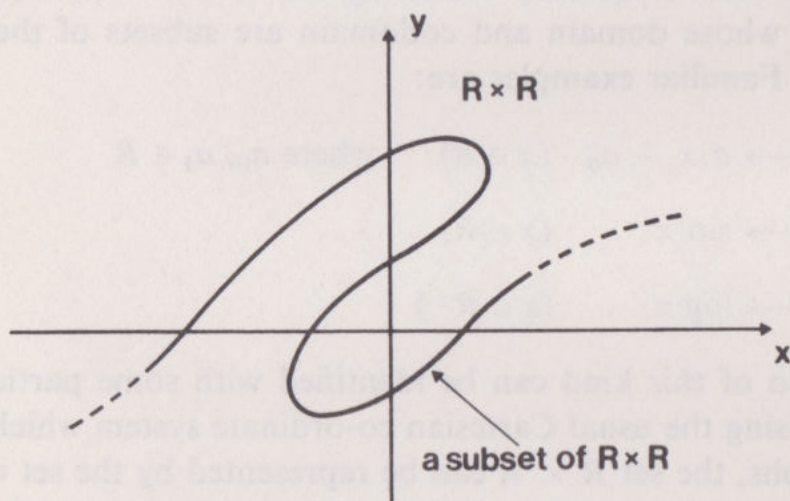
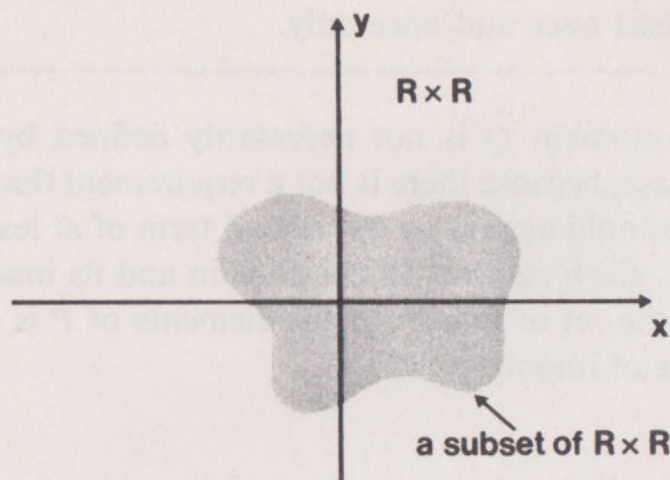
Each function of this kind can be identified with some particular subset of  $R \times R$ . Using the usual Cartesian co-ordinate system which we use for drawing graphs, the set  $R \times R$  can be represented by the set of all points in a plane. Any subset of  $R \times R$  then corresponds to a set of points in the plane.

Thus the subset of  $R \times R$  with which we identify, for example, the function  $x \longmapsto \sin x$ , is represented by





Some other possible subsets of  $R \times R$  are represented by:



(The dotted ends of the curves are intended to mean that the curves extend indefinitely in the final direction indicated.)



*Exercise 1*

Why do *neither* of the first two of the three subsets of  $R \times R$  depicted immediately above define a function  $f: R \longrightarrow R$ ?

*Exercise 2*

For the sets

$$A = \{a, b, c\}$$

$$B = \{1, 2, 3, 4\}$$

write out the Cartesian products  $A \times B$  and  $B \times A$ .

**1.5 Additional Exercises***Exercise 1*

If  $m$  is a mapping which maps a set  $A$  to a set  $B$ , and if  $a \in A$  and  $b \in B$ , which of the following statements are true and which false?

- (i) If  $m: a \longmapsto b$ , then  $m(a) = b$ .
- (ii) If  $m: a \longmapsto b$ , then  $m(a)$  is the image of  $b$ .
- (iii) If  $m: a \longmapsto b$ , then  $m(a) \in B$ .

*Exercise 2*

Say which of the following statements are true and which are false:

- (i) The statement  $f: x \longmapsto x^2 + 1$  ( $x \in R$ ) implies that  $f(2) = 5$
- (ii) The statement  $f: x \longmapsto x^2 + 1$  implies that  $f(1) = 2$  (Be careful!)
- (iii) The statement  $f(2) = 5$  implies that  $f: x \longmapsto x^2 + 1$  ( $x \in R$ )
- (iv) The statement  $f: x \longmapsto 2x + 6$  ( $x \in R^+$ ) implies that  $f(-10) = -14$
- (v) The statement  $f: x \longmapsto 2x + 6$  ( $x \in R$ ) implies that  $f: t \longmapsto 2t + 6$  where  $t$  is any real number
- (vi) The statements  $f: x \longmapsto 6x - 1$  ( $x \in R$ ) and  $f: t \longmapsto 6t - 1$  ( $t \in R$ ) are equivalent
- (vii) The statements  $f: x \longmapsto 4x^2 + 1$  ( $x \in R$ ) and  $f: t \longmapsto 4t^2 + 1$  ( $t \in R^+$ ) are equivalent

*Exercise 3*

State which of the following sets define mappings from the set  $A$  to the set  $B$ , where

$$A = \{a, b, c\}$$

and

$$B = \{1, 2, 3\}$$

- (i)  $\{(a, 1), (b, 2), (c, 3)\}$
- (ii)  $\{(a, 1), (a, 2), (a, 3)\};$
- (iii)  $\{(a, 1), (a, 3), (b, 2), (b, 1), (c, 3)\};$
- (iv)  $\{(a, 1), (b, 1), (c, 2)\}.$

Which of these mappings are functions?

*Exercise 4*

Calculate reverse mappings or inverse functions for the functions defined as follows:

- (i)  $f: x \mapsto 7x - 1 \quad (x \in R)$
- (ii)  $f: x \mapsto 4x^2 + 3 \quad (x \in R)$

## 1.6 Answers to Exercises

### Section 1.1

*Exercise 1*

- (i) (a), (e), (f), (g) are correct.
- (ii) None of the statements is correct.

The fact that  $A$  and  $B$  have the same number of elements is not enough to give equality. For two sets to be equal, not only must they have the same number of elements but these elements must be the same.

### Section 1.2

*Exercise 1*

- (i) 75
- (ii) 51
- (iii) 1944
- (iv) 20
- (v) 240



- (vi) 10
- (vii) {75, 120, 175}

### Exercise 2

- (i) FALSE
- (ii) FALSE
- (iii) TRUE  
By definition, the codomain contains all the images of the elements of the domain.
- (iv) FALSE  
The codomain may contain elements which are not images under the mapping.
- (v) TRUE  
The list satisfies the requirements of a “mapping rule”. It assigns at least one element of  $B$  to *each* element of  $A$ .
- (vi) FALSE  
The mapping is not a function because  $\alpha$  does not have a *unique* element as its image.
- (vii) FALSE  
No element is assigned to  $\gamma$ .

### Exercise 3

- (i) (a)  $R^+$       (b) {3, 5, 7}      (c) The set of real numbers greater than 1
- (ii) (a)  $R$       (b) {−1, 2, 7}      (c) The set of real numbers greater than or equal to −2
- (iii) (a)  $R$       (b) 3      (c) 3

This last function is an example of a constant function. A **constant function** is one for which the image of every element of the domain is the same.

### Exercise 4

- (i) Many-one, e.g.  $1 \mapsto 5$  and  $-1 \mapsto 5$ .
- (ii) One-one.
- (iii) Many-one, e.g.  $0 \mapsto 0$ ,  $\pi \mapsto 0$ , etc.
- (iv) One-many. (Had the domain been  $R$  instead of  $R^+$ , the mapping would have been many-many.)
- (v) Many-many, e.g.  $1 \mapsto \{2, -2\}$ ,  $-1 \mapsto \{2, -2\}$ .

## Section 1.3

## Exercise 1

- (i)  $g + f: x \mapsto 6x + 6x^2 \quad (x \in [-1, 1])$   
 (ii)  $g \div f: x \mapsto 1/x \quad (x \in [-1, 1] \text{ and } x \neq 0)$   
 (iii)  $f \div g: x \mapsto x \quad (x \in [-1, 1] \text{ and } x \neq 0)$

(Although  $f(x)/g(x) = x$ , this is only true where we can actually perform the division. Therefore mathematically, and by definition, we must exclude  $x = 0$ .)

- (iv)  $f \times g: x \mapsto 36x^3 \quad (x \in [-1, 1])$

## Exercise 2

- (i) (a)  $f \circ g: x \mapsto x^2 - 1 \quad (x \in R)$   
 (b)  $g \circ f: x \mapsto (x - 1)^2 \quad (x \in R)$   
 (ii) It ought to be  $g \circ f$ , but in practice you would often get a different result going to German via French rather than directly from English.

## Exercise 3

- (i) NO. We can form  $g \circ f$  only if the set of all images of the domain of  $f$  is a subset of (or equal to) the domain of  $g$ . Consider, for example,

$f: \text{person} \mapsto \text{colour of his eyes (domain the set of all people)}$

$$g: x \mapsto x^2 \quad (x \in R)$$

Then we cannot form  $g \circ f$ , since the square of a colour is not defined.

- (ii) NO. We now have restrictions on the set of images under  $g$  and on the domain of  $f$ . Consider, for example,

$$f: x \mapsto \sqrt{x} \quad (x \in R^+)$$

$$g: x \mapsto x + 3 \quad (x \in R)$$

Then

$$g \circ f: x \mapsto \sqrt{x} + 3 \quad (x \in R^+) \text{ is satisfactory}$$

but

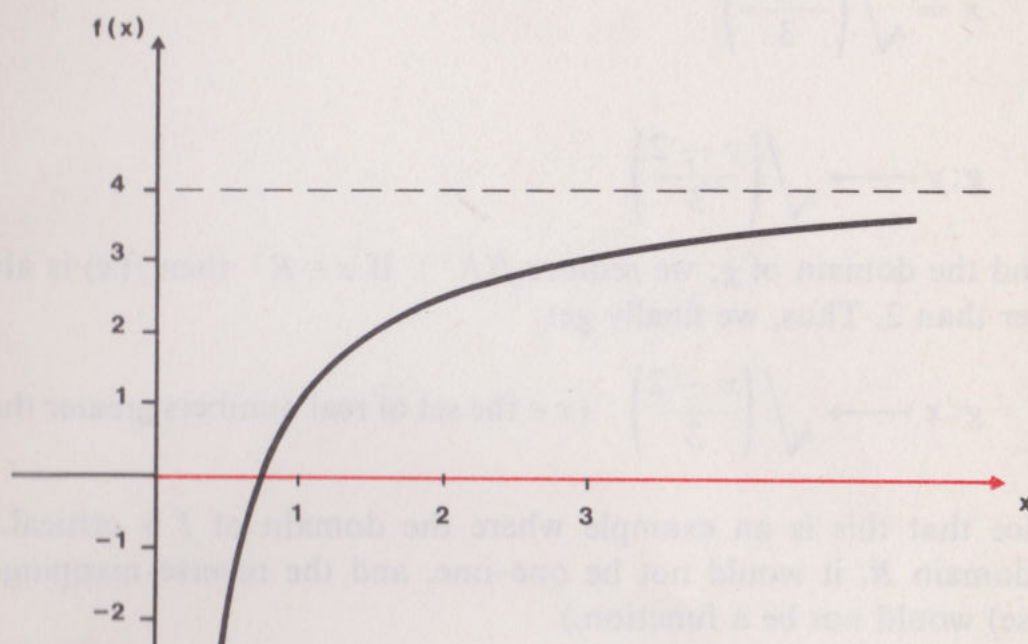
$$f \circ g: x \mapsto \sqrt{(x + 3)} \text{ is undefined for } x < -3.$$



## Exercise 4

$$g: x \mapsto \frac{3}{4-x} \quad (x \in \text{the set of real numbers less than } 4)$$

Remember that the domain is  $f(\mathbb{R}^+)$ . Since  $x$  is positive,  $4 - \frac{3}{x}$  is always less than 4. The graph of  $f$  shows quite clearly that the set of images is the set of all real numbers less than 4.



## Exercise 5

- (i) YES. We have actually defined the inverse function of  $f$  only when  $f$  is one-one. As we have hinted in the text, we could have defined an inverse function of a one-many mapping  $f$ ; then of course,  $f$  is not a function.
- (ii) Both  $g \circ f$  and  $f \circ g$  are given by the formula  $x \mapsto x$ . But the domain of  $g \circ f$  is that of  $f$ , and the domain of  $f \circ g$  is that of  $g$ . So  $f \circ g = g \circ f$  only if  $f$  and  $g$  have the same domain.

For instance, in the previous exercise you can check that the formula for both  $g \circ f$  and  $f \circ g$  is  $x \mapsto x$ . But the domain of  $g \circ f$  is the domain of  $f$ , i.e.  $\mathbb{R}^+$ , whereas the domain of  $f \circ g$  is the domain of  $g$ , i.e. the set of real numbers less than 4.

- (iii) NO. If  $f: x \mapsto x$ , then  $g: x \mapsto x$ , so  $g(x) \neq \frac{1}{x}$  unless the domain of  $f$  is  $\{1\}$ .

## Exercise 6

Rearranging the equation we get

$$x = \pm \sqrt{\left(\frac{y-2}{3}\right)}$$

Since  $f$  is one-one we must choose the signs correctly and get a single value for  $x$  in terms of  $y$ . Since  $x \in \mathbb{R}^+$ , we choose the positive sign and so

$$x = \sqrt{\left(\frac{y-2}{3}\right)}$$

or

$$g: y \longmapsto \sqrt{\left(\frac{y-2}{3}\right)}$$

To find the domain of  $g$ , we require  $f(\mathbb{R}^+)$ . If  $x \in \mathbb{R}^+$  then  $f(x)$  is always greater than 2. Thus, we finally get

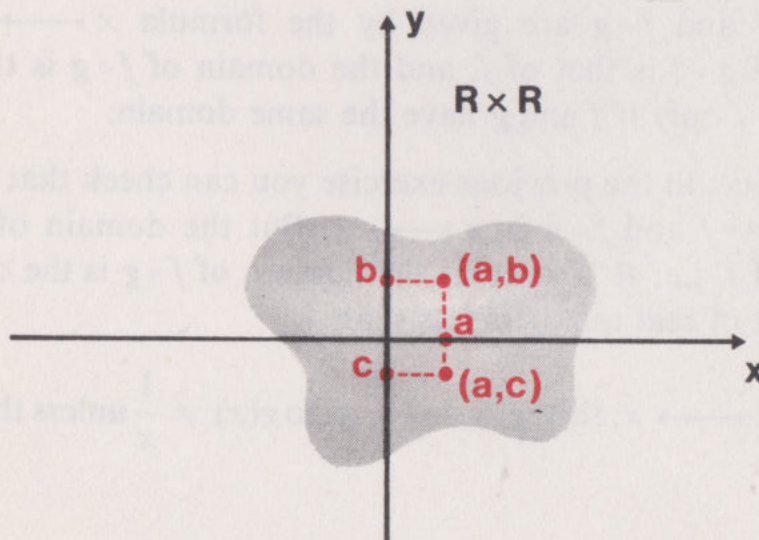
$$g: x \longmapsto \sqrt{\left(\frac{x-2}{3}\right)} \quad (x \in \text{the set of real numbers greater than 2})$$

(Notice that this is an example where the domain of  $f$  is critical. If  $f$  had domain  $\mathbb{R}$ , it would not be one-one, and the reverse mapping (*not* inverse) would not be a function.)

## Section 1.4

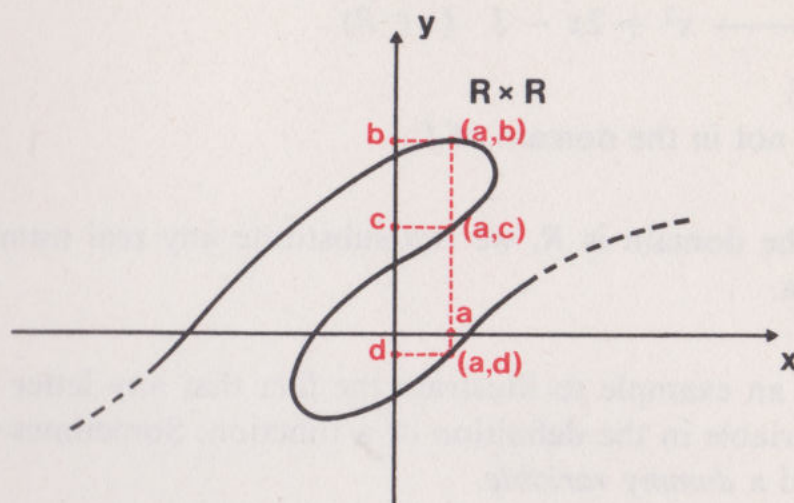
## Exercise 1

No *area* can define a function  $\mathbb{R} \rightarrow \mathbb{R}$ . If the ordered pair  $(a, b)$  belongs to the area then there are other ordered pairs  $(a, c)$  with  $c \neq b$  also belonging to the area, and  $a$  thus has more than one element in its image.





Similarly, in the case of the second subset illustrated, there will be elements such as  $a$  having more than one element in its image because of the way in which the curve “folds back” on itself.



### Exercise 2

$$A \times B = \{(a, 1), (a, 2), (a, 3), (a, 4), (b, 1), (b, 2), (b, 3), (b, 4), (c, 1), (c, 2), (c, 3), (c, 4)\}$$

$$B \times A = \{(1, a), (1, b), (1, c), (2, a), (2, b), (2, c), (3, a), (3, b), (3, c), (4, a), (4, b), (4, c)\}$$

## Section 1.5

### Exercise 1

- (i) TRUE
- (ii) FALSE
- (iii) TRUE ( $m(a)$  stands for the image of  $a$  under  $m$ . The image belongs to the set  $B$ , and so  $m(a) \in B$ ).

### Exercise 2

- (i) TRUE

The statement defines  $f$  completely; 2 is in the domain and so we can substitute in the formula, and  $2 \times 2 + 1 = 5$ .

- (ii) FALSE

The statement does not define  $f$  because the domain is not specified and so we do not know whether we are allowed to substitute the number 1 into the formula.

(iii) FALSE

$f$  could be one of any number of functions which include 2 in the domain and which map 2 to 5, e.g.

$$f: x \mapsto x^2 + 2x - 3 \quad (x \in R)$$

(iv) FALSE

$-10$  is not in the domain of  $f$ .

(v) TRUE

Since the domain is  $R$ , we *can* substitute any real number into the formula.

(vi) TRUE

This is an example to illustrate the fact that any letter can be used as a variable in the definition of a function. Sometimes such a letter is called a *dummy variable*.

(vii) FALSE

The domains are not the same.

### Exercise 3

(i), (iii), and (iv) define mappings; (ii) does not, because no images are assigned to the elements  $b$  and  $c$ .

The mappings (i) and (iv) define functions.

The mapping (iii) does not define a function because, for instance, the element  $a$  does not have a unique element as its image.

### Exercise 4

$$(i) \quad g: x \mapsto \frac{x+1}{7} \quad (x \in R)$$

$$(ii) \quad g: x \mapsto \left\{ \sqrt{\left(\frac{x-3}{4}\right)}, -\sqrt{\left(\frac{x-3}{4}\right)} \right\} \quad (x \in \text{the set of real numbers greater than or equal to } 3)$$



## CHAPTER 2 OPERATIONS AND RELATIONS

### 2.0 Introduction

In this chapter we introduce the concepts of *operation* and *relation*. These two concepts are connected, as we shall see, and both bear an important relation to the concept of *function*.

In the course of our discussion of operations we introduce *union* and *intersection* as binary operations on a set of sets, and *complementation* as a unary operation on a set of sets.

Two particular kinds of relation are of special importance because they provide mathematical models of the process of *sorting* and *ordering*. Because sorting and ordering are such common everyday phenomena, mathematicians ask themselves whether such processes have intrinsic features which can be abstracted in order that mathematical models of the processes can be made. To discover the answer to this question it is first necessary to describe sorting and ordering situations in mathematical language. This leads us to a discussion of *equivalence relations*, which model the sorting process, and *order relations*, which model ordering.

### 2.1 Binary Operations

We will begin this section by investigating operations such as addition and multiplication, by which we combine two numbers to produce another number. Such operations are called *binary operations*.

The word “binary” arises because we combine *two* numbers. There are operations which act on one object and operations which operate on three (or more) objects. We shall discuss these other operations briefly in section 2.2, but our main interest is in *binary* operations.

#### *Example 1*

We are all familiar with  $+$  representing the operation of addition in the set of real numbers  $R$ . For example, we have

$$3 + 5 = 8$$

We could rewrite this in our mapping notation as

$$+ : (3, 5) \longmapsto 8$$

This is not a particularly helpful way of writing down an addition sum,



but it does illustrate the fact that the binary operation of addition on  $R$  defines a function with domain  $R \times R$  and codomain  $R$ . (Note that this domain permits us to combine an element of  $R$  with itself.)

As usual, we do not want to restrict ourselves to numbers, so we adopt the following definition of a binary operation:

A **binary operation**,  $\circ$ , on a set  $A$  is a rule which assigns to *each* ordered pair  $(a_1, a_2) \in A \times A$  a uniquely defined element  $b$ .

This is equivalent to saying that a binary operation on  $A$  is a function with domain  $A \times A$ , and codomain some set  $B$ .

We write

$$\begin{array}{c} a_1 \circ a_2 = b \\ \downarrow \\ \text{read} \\ \text{as} \\ \text{"circle"} \end{array}$$

If  $a_1 \circ a_2$  belongs to  $A$  for all  $a_1, a_2 \in A$ , then we say that the binary operation is **closed**.

(More precisely, we should say that  $A$  is *closed for*  $\circ$ , since an operation cannot exist without a set. We shall, however, abbreviate.) A codomain of the function defined by a closed binary operation on  $A$  is therefore  $A$  itself.

### Example 2

Consider the binary operation of addition,  $+$ , on the set  $\{0, 1, 2, 3, 4\}$ . This is *not* closed because there are several pairs of elements of the set which can be added to give a result which is not a member of the original set, e.g.  $2 + 4$ . We can, however, devise a special kind of "addition" which will be a closed binary operation. One way of doing this is to add first, and then take the remainder on division by 5.

We shall now find that we always end up with an element of our set. Using the symbol  $\oplus_5$  to denote this special sort of addition, we have, for example,

$$2 \oplus_5 4 = 1$$

$$3 \oplus_5 4 = 2$$

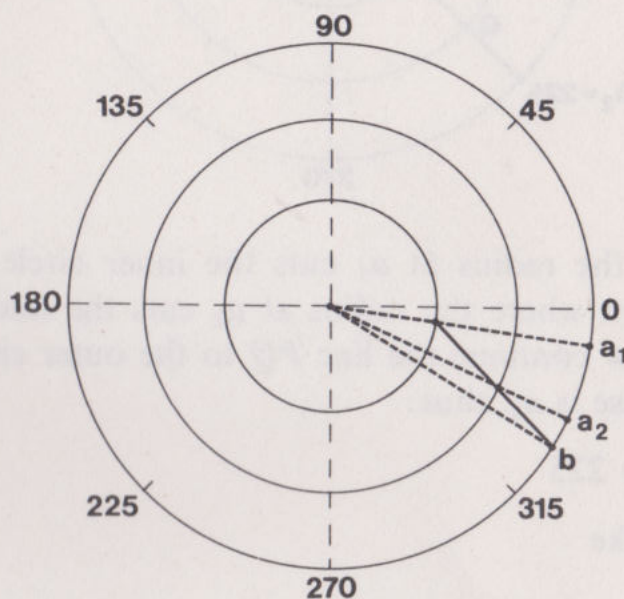


$$4 \oplus_5 4 = 3$$

Thus, " $\oplus_5$ " is a closed binary operation on the set  $\{0, 1, 2, 3, 4\}$ , but " $+$ " is not.

We shall now look at another example in order to bring out a further property possessed by certain binary operations.

### Example 3



In this example the set  $A$  is

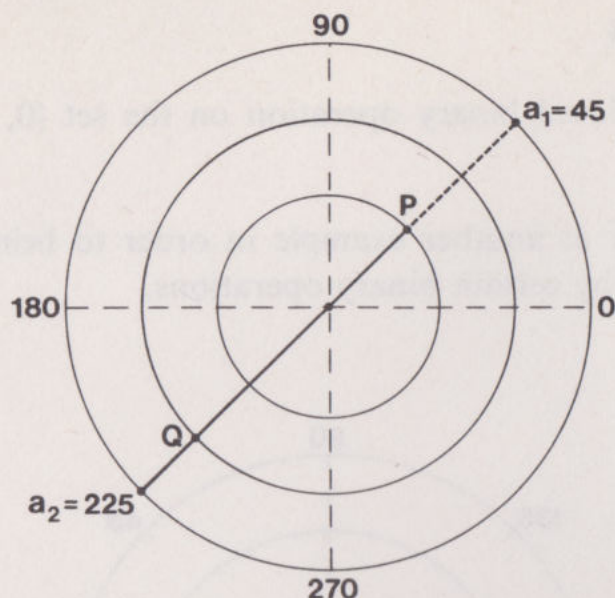
$$\{a : a \in R \text{ and } 0 \leq a < 360\}$$

We illustrate this set as the set of all points on the outer circle of our diagram, and we read off the corresponding numerical value by using the outside scale, as when reading degrees using a protractor. We now give a rule which enables us to combine two elements of  $A$  by an operation  $\circ$  as follows:

Take the point where the radius corresponding to  $a_1$  cuts the inner circle and join it by a straight line to the point where the radius corresponding to  $a_2$  cuts the middle circle. Continue this line to obtain the point  $b$  on the outer circle. The numerical value at  $b$  is defined to be  $a_1 \circ a_2$ .

For example, suppose that we take

$$a_1 = 45, \quad a_2 = 225$$



The point where the radius at  $a_1$  cuts the inner circle is the point  $P$ . Similarly, the point where the radius at  $a_2$  cuts the middle circle is the point  $Q$ . If we now continue the line  $PQ$  to the outer circle we arrive at  $b$ , which in this case is  $a_2$ , thus:

$$45 \circ 225 = 225$$

If, however, we take

$$a_1 = 225$$

$$a_2 = 45$$

then we find that we obtain

$$225 \circ 45 = 45$$

So the *order* in which we take out pair of numbers is important, and for a binary operation  $\circ$  on a set  $A$ ,  $a_1 \circ a_2$  is *not necessarily the same as*  $a_2 \circ a_1$ . If, as in the case of addition (and multiplication) on  $R$ , we have

$$a_1 \circ a_2 = a_2 \circ a_1$$

for *all*  $a_1, a_2 \in A$ , then the binary operation is said to be **commutative**. (This particular adjective is used because we may “commute” (i.e. interchange) the order of the elements.)

### Exercise 1

In each of the following cases determine whether or not the binary operation  $\circ$  on the set  $A$  is *closed*.

- (i)  $a_1 \circ a_2$  is  $a_1 + a_2$ ;

$A$  is the set of all real numbers,  $R$ .



- (ii)  $a_1 \circ a_2$  is  $a_1 - a_2$ ;  
 $A$  is the set of positive integers,  $Z^+$ .
- (iii)  $a_1 \circ a_2$  is  $a_1 \div a_2$ ;  
 $A$  is the set of integers,  $Z$ , excluding zero.
- (iv)  $a_1 \circ a_2$  is the mid-point of the straight line joining  $a_1$  and  $a_2$ ;  
 $A$  is the set of all points on a square piece of paper.
- (v)  $a_1 \circ a_2$  is the mid-point of the straight line joining  $a_1$  and  $a_2$ ;  
 $A$  is the set of all points on a square piece of paper with a circular hole cut out of it.

### Exercise 2

Classify the following list into two categories; those operations on  $A$  which are commutative, and those which are not commutative.

- (i)  $a_2 + a_2$ ;
  - (ii)  $a_1 - a_2$ ;
  - (iii)  $\sin(a_1 + a_2)$ ;
- }  $A$  is the set  $R$
- (iv)  $a_1 \div a_2$ ;  $A$  is the set  $R$ , excluding zero.
  - (v) The mid-point of the straight line joining points  $a_1$  and  $a_2$  on a square piece of paper;  $A$  is the set of all points on the paper.

### Repeated Operations

Although by definition a binary operation  $\circ$  on a set  $A$  combines only two elements, if the result of the operation is also an element of  $A$  (i.e. if  $\circ$  is *closed* on  $A$ ), then we can combine the result with a further element. Let us suppose then that for all  $a_1, a_2 \in A$ , we have an operation  $\circ$  which gives us  $a_1 \circ a_2$  in  $A$ , and let us now combine this result with  $a_3$ , where  $a_3 \in A$ . We denote the total result by

$$(a_1 \circ a_2) \circ a_3$$

The brackets enclose the elements that are combined first.

We have already seen that the order of the elements is important unless the operation is commutative, but does it matter if we combine  $a_1$  and  $a_2$  first, or  $a_2$  and  $a_3$  first? What we are asking in effect is:

$$\text{does } (a_1 \circ a_2) \circ a_3 \text{ equal } a_1 \circ (a_2 \circ a_3)?$$

The order of the terms in each expression is the same, but the way in which the terms are grouped is different.



*Exercise 3*

Of the following statements, which are true and which are false?

- (i)  $x + (y + z) = (x + y) + z$ , where  $x, y, z \in R$ .
- (ii)  $x - (y - z) = (x - y) - z$ , where  $x, y, z \in R$ .
- (iii)  $(x^y)^z = x^{(y^z)}$ , where  $x, y, z \in Z^+$ .  
(Here  $x \circ y = x^y$ , hence  $(x \circ y) \circ z = (x^y)^z$ , etc.)
- (iv)  $x \div (y \div z) = (x \div y) \div z$ , where  $x, y, z \in R^+$ .
- (v)  $(x \times y) \times z = x \times (y \times z)$ , where  $x, y, z \in R$ .
- (vi)  $P \circ (Q \circ R) = (P \circ Q) \circ R$ , where  $P, Q, R$  are points in a plane and  $P \circ Q$  is the mid-point of the straight line joining  $P$  and  $Q$ .
- (vii)  $x \circ (y \circ z) = (x \circ y) \circ z$ , where  $x, y, z \in R$ , and  $\circ$  is the operation of addition following by rounding-off to three significant figures.

When the closed binary operation is such that

$$(a_1 \circ a_2) \circ a_3 = a_1 \circ (a_2 \circ a_3)$$

for *all* elements of  $A$ , we say that the operation is **associative**. This particular adjective is used to denote this property because we may “associate” either  $a_1$  and  $a_2$  (as on the left-hand side) or  $a_2$  and  $a_3$  (as on the right-hand side).

*Exercise 4*

Why was it necessary to require that the binary operation on  $A$  be *closed* when we discussed the idea of associativity?

**Two Binary Operations**

Suppose now that we have a set with two closed binary operations. As an example we shall again take the set of all real numbers  $R$ , and we shall consider the operation of multiplication as well as addition. We start with three elements of  $R$ .

We have, for example,

$$2 \times (3 + 5)$$

We know that this is equal to

$$(2 \times 3) + (2 \times 5)$$

since in each case we obtain the number 16. In fact, no matter which three



numbers we take, we always find that

$$x \times (y + z) = (x \times y) + (x \times z)$$

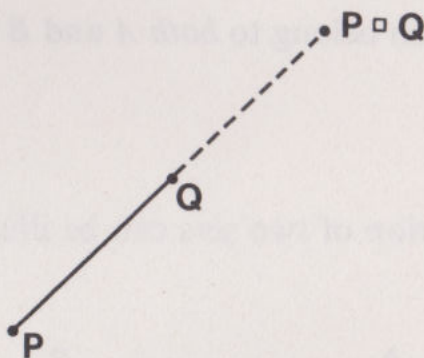
whenever  $x, y, z \in R$ .

This property of  $\times$  and  $+$  on  $R$  is a very useful one: for instance, it helps to simplify calculations, and without it the factorization of algebraic expressions would not be permissible. This leads us to ask of other pairs of binary operations whether they will obey a similar law. This is precisely the question asked in the next exercise for some particular pairs of binary operations.

### Exercise 5

Of the following statements, which are true and which are false?

- (i)  $x + (y \times z) = (x + y) \times (x + z)$ , where  $x, y, z \in R$ .
- (ii)  $x + (y - z) = (x + y) - (x + z)$ , where  $x, y, z \in \mathbb{Z}$ .
- (iii)  $x \times (y - z) = (x \times y) - (x \times z)$ , where  $x, y, z \in R$ .
- (iv)  $(y - z) \times x = (y \times x) - (z \times x)$ , where  $x, y, z \in R$ .
- (v)  $P \square (Q \circ R) = (P \square Q) \circ (P \square R)$ , where  $P, Q, R$  are points in a plane,  $Q \circ R$  is the mid-point of the straight line  $QR$ , and  $P \square Q$  (read as  $P$  "square"  $Q$ ) is the point on the line  $PQ$  extended so that the distance from  $P$  to  $Q$  is the same as the distance from  $Q$  to  $P \square Q$ .



- (vi)  $(Q \circ R) \square P = (Q \square P) \circ (R \square P)$ , where  $P, Q, R$ ,  $\circ, \square$  are the same as for (v).
- (vii)  $(x + y) \div z = (x \div z) + (y \div z)$ , where  $x, y, z \in R^+$ .
- (viii)  $z \div (x + y) = (z \div x) + (z \div y)$ , where  $x, y, z \in R^+$ .



When two closed binary operations,  $\square$  and  $\circ$ , on a set  $A$  have the property that

$$a_1 \square (a_2 \circ a_3) = (a_1 \square a_2) \circ (a_1 \square a_3)$$

for *all* elements of  $A$ , we say that the operation  $\square$  is **left-distributive** over the operation  $\circ$ .

So, multiplication is left-distributive over addition on  $R$ . ("Left" is used because  $\square$  stands on the left (once) before being distributed (twice) over  $\circ$ .)

Similarly 
$$(a_1 \circ a_2) \square a_3 = (a_1 \square a_3) \circ (a_2 \square a_3)$$

is a definition of **right-distributivity** of  $\square$  over  $\circ$ .

If the operation  $\square$  is commutative, then one can be inferred from the other as in (iii) and (iv) of Exercise 5.

On the other hand, parts (vii) and (viii) of this exercise show that division is right-distributive but not left-distributive over addition.

**We shall now define the single term distributive to mean left- and right-distributive.**

### Union and Intersection as Binary Operations

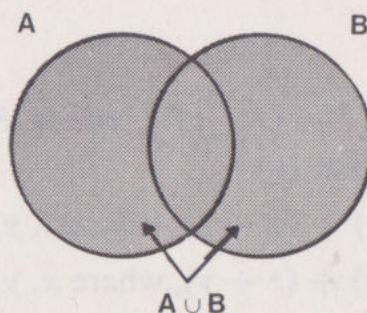
The set of elements which belong to *either or both* of two sets  $A$  and  $B$  is called the **union** of  $A$  and  $B$ , written

$$A \cup B.$$

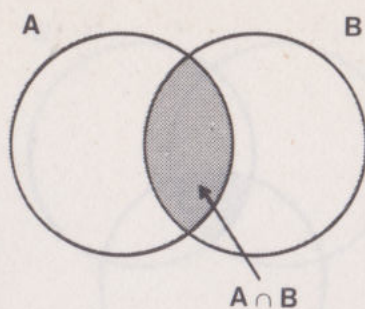
The set of elements which belong to *both*  $A$  and  $B$  is called the **intersection** of  $A$  and  $B$ , written

$$A \cap B.$$

The union and intersection of two sets can be illustrated by the following two diagrams:

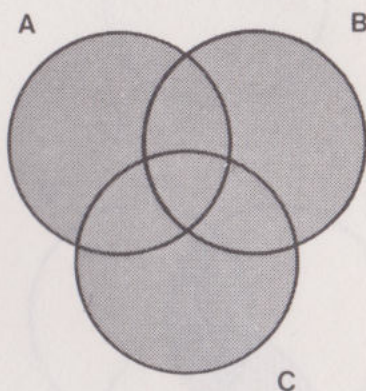
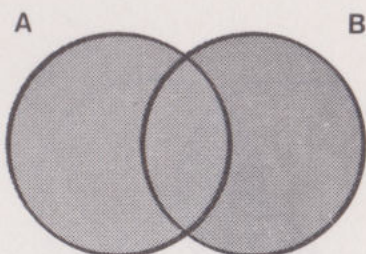






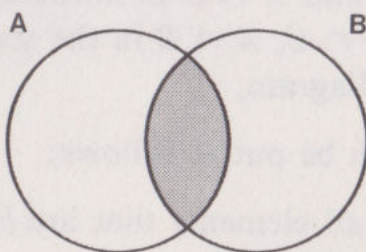
If  $A$  and  $B$  are subsets of some set  $X$ , then so also are  $A \cap B$  and  $A \cup B$ ; in other words,  $\cap$  and  $\cup$  can be regarded as *binary operations* on the set of all subsets of  $X$ , analogous to the binary operations  $\cdot$  and  $+$  on  $R$ . How far does this analogy extend? We shall show below that it extends quite far, by setting out some properties of associativity, commutativity and distributivity.

$\cup$  is both commutative and associative.

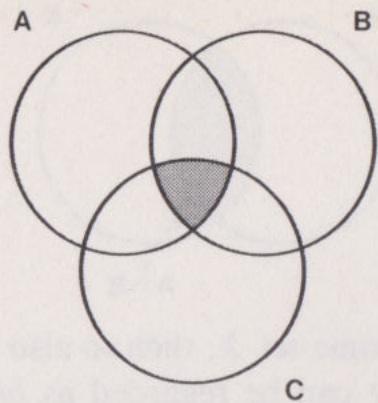


(The first diagram illustrates both  $A \cup B$  and  $B \cup A$ . The second diagram illustrates both  $A \cup (B \cup C)$  and  $(A \cup B) \cup C$ . The union of a collection of sets is simply the set of all elements *which belong to at least one* of the sets, and is independent of the order in which the sets are written down.)

$\cap$  is both commutative and associative.

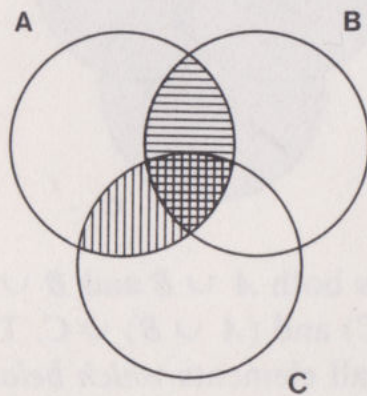
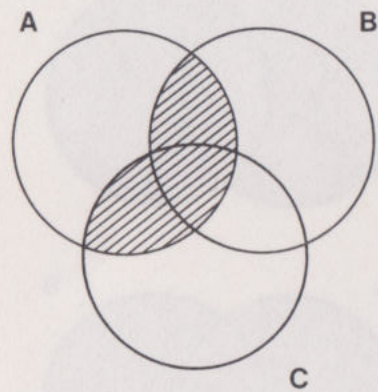






(The intersection of a collection of sets is simply the set of all elements *which are common to all* the sets, and is again independent of the order in which the sets are written down.)

$\cap$  is distributive over  $\cup$ .



The first diagram represents  $A \cap (B \cup C)$ , and the second diagram shows  $A \cap C$  in vertical shading and  $A \cap B$  in horizontal shading. Clearly, the union of the shaded sets  $A \cap C$ ,  $A \cap B$  in the second diagram is equal to the shaded set in the first diagram.

In words, the argument can be put as follows:

$A \cap (B \cup C)$  is the set of all elements that are *both* in  $A$  and in  $B \cup C$ .



That is, they are in  $A$  anyway, and also in  $B$  or  $C$  or both  $B$  and  $C$ . Thus, they are in  $A$  and in  $B$ , or they are in  $A$  and in  $C$ , or in  $A$  and in  $B$  and  $C$ . Therefore, they constitute the set  $(A \cap B) \cup (A \cap C)$ .

### Exercise 6

Use both diagrams and verbal argument to establish the following property.

$\cup$  is distributive over  $\cap$ .

The two distributive properties of the operations  $\cap$  and  $\cup$  for sets:

$$\left. \begin{aligned} A \cap (B \cup C) &= (A \cap B) \cup (A \cap C) \\ A \cup (B \cap C) &= (A \cup B) \cap (A \cup C) \end{aligned} \right\} \text{ for all } A, B, C \in X$$

remind us of the distributive property of  $\cdot$  over  $+$  for real numbers:

$$a \cdot (b + c) = (a \cdot b) + (a \cdot c) \quad \text{for all } a, b, c \in R$$

But note that the operations  $\cup$  and  $\cap$  for sets are not completely analogous to the operations  $\cdot$  and  $+$  for real numbers, for

$$a + (b \cdot c) \neq (a + b) \cdot (a + c) \quad \text{for all } a, b, c \in R$$

i.e.  $+$  is *not* distributive over  $\cdot$ , but  $\cap$  and  $\cup$  are each distributive over the other.

## 2.2 N-ary Operations

In section 2.1, we saw that a binary operation on  $A$  defines a function with domain  $A \times A$  and vice versa. We ought not to be surprised at this very close relation between a binary operation and a function because, after all, the idea of a *rule* is present in both concepts.

We have so far concentrated our interest on *binary* operations because the idea of an operation as such comes to mind most naturally when we consider combining two elements of a given set. But the close connection between a binary operation on  $A$  and a function with domain  $A \times A$  leads us to re-examine functions briefly.

### Example 1

Consider the function

$$x \longmapsto x^2 \quad (x \in R)$$



The rule which tells us how to obtain the image of any given element of  $R$  can be regarded as an *operation* on  $R$ . In this case the operation is

square it

Because here we are operating on *only one* element of  $R$  at a time, we call such an operation a **unary operation**.

Example 1 illustrates a function  $f:A \longrightarrow B$ , where the image set is a subset of the domain  $A$ , and we can therefore interpret this function as a *closed* unary operation.

We can also ask ourselves about operations on a set which combine three or more elements to give one element of the same or a different set. Again there is the same close connection with the idea of a function, and we can construct a simple table to illustrate this.

Operation $\circ$ on a set $A$ acts on:	Operation is called	Corresponding function
single element $a_1$	UNARY	$f:A \longrightarrow B$
ordered pair $(a_1, a_2)$	BINARY	$f:A \times A \longrightarrow B$
ordered triple $(a_1, a_2, a_3)$	TERNARY	$f:A \times A \times A \longrightarrow B$
.....	.....	.....
ordered $n$ -tuple $(a_1, a_2, \dots, a_n)$	$N$ -ARY	$f:\underbrace{A \times A \times \dots \times A}_{n \text{ times}} \longrightarrow B$

In the table we have introduced the notation  $A \times A \times A$  to mean the set

$$\{(a_1, a_2, a_3): a_1, a_2, a_3 \in A\}$$

i.e. the set of all ordered triples that can be formed from elements of  $A$ . This is the notation which is usual in the mathematical literature. The notation is extended in the obvious way as we go down the table.

Whether we choose to talk about an operation or its corresponding function is purely a matter of context. Sometimes it is more natural to



speak of such and such an operation, sometimes of such and such a function.

### Example 2

Consider the function

$$(x, y, z) \longmapsto xyz \quad ((x, y, z) \in R \times R \times R)$$

If we consider this as an *operation on  $R$* , then it is a *ternary operation*, since it combines by multiplication three elements of  $R$ . However, we can also consider it as an *operation on  $R \times R \times R$* , in which case it is a *unary operation*, since it operates on only one element of  $R \times R \times R$  at a time.

### Exercise 1

How would you define the set  $A$  so that the function

$$(x, y) \longmapsto x^3y^2 \quad ((x, y) \in R \times R)$$

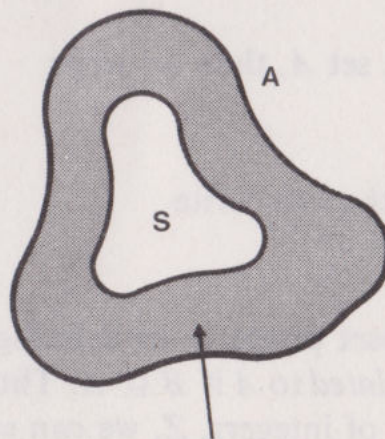
corresponds to

- (i) a binary
- (ii) a unary

operation on  $A$ ?

Which, if either, of the operations would be closed?

Whenever we define a subset,  $S$ , of a given set  $A$ , we automatically define another set, the set of all those elements of  $A$  which do *not* belong to  $S$ . In terms of the diagrams we have been using, and representing a proper subset of the set  $A$  by a region contained entirely within the region representing  $A$ , this set is indicated as follows:



The set of all elements of  $A$   
which do not belong to  $S$



This set is called the **complement** of  $S$  with respect to  $A$ . If, in any argument involving complements, all the complements are with respect to the same set, then we usually denote the complement of a set  $S$  by  $S'$ .

The operation of taking the complement of subsets of a given set  $A$  is a unary operation on the set of all subsets of  $A$ . The corresponding function maps each subset  $S$  to its complement  $S'$  with respect to  $A$ , and has domain and codomain the set of all subsets of  $A$ .

### Exercise 2

All the following sets are to be considered as subsets of the set of all people. If  $A$  is the set of all males,  $B$  is the set of all people who speak English, describe in words the following sets:

- (i)  $A'$
- (ii)  $B'$
- (iii)  $(A')'$
- (iv)  $(A \cap B)'$
- (v)  $A' \cup B'$

## 2.3 What is a Relation?

When we talk of two people being related, we are using the term in a pretty imprecise way. Some people may stop at cousins when considering relations ; others may include a brother-in-law's aunt. In mathematics the word retains much of its usual meaning, but we have, as always, to define our words precisely. So we shall start by looking at mathematical examples and then proceed to definitions.

### Example 1

If a set  $B$  is a subset of a set  $A$ , then we write

$$B \subseteq A,$$

and if it is not a subset, then we write

$$B \not\subseteq A.$$

We can think of the subset property as defining a *relation* between sets: that is, we say that  $B$  is *related* to  $A$  if  $B \subseteq A$ . Thus, for the set of positive integers,  $Z^+$ , and the set of integers,  $Z$ , we can say

$$Z^+ \subseteq Z,$$



that is,  $Z^+$  is *related* to  $Z$ , but

$$Z \not\subseteq Z^+,$$

that is,  $Z$  is not *related* to  $Z^+$ .

Here we are dealing with a particular relation, which we may put into words as “is a subset of” (cf. “is a brother of”). However, such a phrase *on its own* is not sufficient to specify a relation, because we need also to state the set or sets whose elements are being compared. For example, starting with a specified person, Fred Jones, the relationship “is the father of” might give completely different answers if it were defined on the set of all people to those obtained if it were defined on the set of all males.

### Example 2

Earlier in this chapter we defined a *binary operation* on a set. For example, the binary operation of multiplication on  $Z$  gives us

$$7 \times 3 = 21.$$

We shall say that the integer 21 is *related* to the ordered pair of integers  $(7, 3)$  by a relation which we may put into words as “is the product of”. Notice again that we must specify the sets from which we are comparing elements, in this case the set  $Z$  and the set  $Z \times Z$  (the Cartesian product of  $Z$  with itself).

### Example 3

In Chapter 1 we defined a *mapping*, and in this case we could say that an element of the codomain is *related* to an element of which it is an image by a relation which we can put into words as “is an image of” (under the appropriate mapping). Again, for the relation to be properly defined we must know the sets whose elements are being compared; these could be the codomain and domain.

We have looked at these three examples because we want to highlight that a *relation* is something *very general indeed*, and because of this generality, relations play an important role in mathematics.

Before formally defining a *relation* we shall consider further examples.

### Example 4

Consider the following two sets of names of men and of women respectively:



$$A = \{\text{Jim, Fred, Tom, Arthur}\},$$

$$B = \{\text{Mary, Anne, Sarah, Karen, Jane}\}$$

We can compare elements from these sets by, for example, considering which (if any) of the men are married to one of the women. If we start with a member of set  $A$  and compare this member with the members of set  $B$  under the relationship which we may express in words as “is the husband of”, we can make a list of statements, such as this:

Jim is not the husband of Mary,  
 Jim is the husband of Anne,  
 Jim is not the husband of Sarah,  
 ...

We can then consider Fred’s relationship with the women and obtain:

Fred is not the husband of Mary,  
 Fred is not the husband of Anne,  
 ...

We shall perhaps find that Fred is not the husband of any of the women named in set  $B$ . We can continue the list until we have exhausted all the possibilities. We shall obtain *two sets of ordered pairs*; one of pairs for which the given relationship holds, and one of pairs for which the relationship does not hold.

We could, for example, obtain a set of married couples:

$\{(\text{Jim, Anne}), (\text{Tom, Mary}), (\text{Arthur, Jane})\}$ ,

and a set of unmarried pairs:

$\{(\text{Jim, Mary}), (\text{Jim, Sarah}), (\text{Jim, Karen}),$   
 $(\text{Jim, Jane}), (\text{Fred, Mary}), \dots, (\text{Arthur, Karen})\}$

At a first glance, it may appear that in our example all we are doing is describing some particular kind of mapping. In fact a relation is *more general than a mapping*.

### Exercise 1

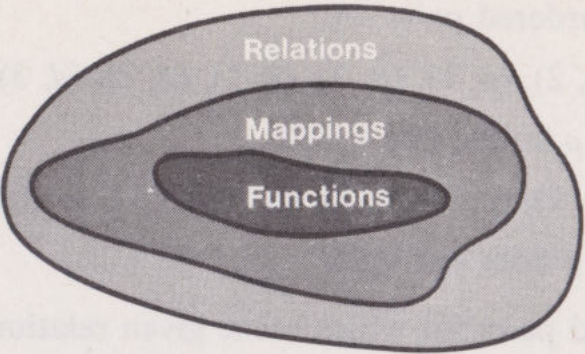
Why does the set of ordered pairs

$\{(\text{Jim, Anne}), (\text{Tom, Mary}), (\text{Arthur, Jane})\}$

not define a mapping with domain the set  $A$  and codomain the set  $B$ ?

We see that *a mapping can be expressed as a relation, but a relation is not necessarily a mapping*.

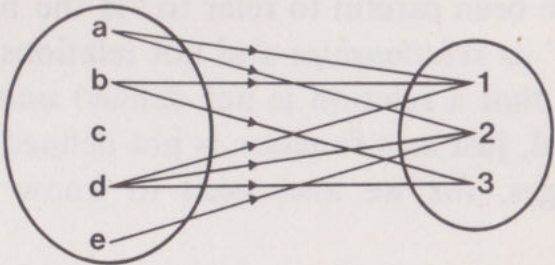




Next we look at a more abstract example.

Example 5

Consider the diagram:



We can again compile two sets of ordered pairs by considering which letters are linked to which numbers. The relationship can be expressed in words as “is linked by a line to”, but we shall adopt a more general notation and write

$$a \rho 1$$

to show that *a* is related to 1, and

$$a \not\rho 3$$

to show that *a* is not related to 3.

We use the Greek letter  $\rho$  (rho) instead of the more natural  $R$ , because we are already using  $R$  for the set of real numbers. We shall refer to *the relation*  $\rho$ .

We see from the diagram that

$$a \rho 1$$

$$a \rho 2$$

$$a \not\rho 3$$

$$b \rho 1, \text{ etc.}$$



Our two sets of ordered pairs are:

$$\{(a, 1), (a, 2), (b, 1), (b, 3), (d, 1), (d, 2), (d, 3), (e, 2)\}$$

the set for which  $\rho$  holds, and

$$\{(a, 3), (b, 2), (c, 1), (c, 2), (c, 3), (e, 1), (e, 3)\},$$

the set for which  $\rho$  does not hold.

The set of ordered pairs for which some given relationship holds is called the **solution set** of the relation concerned. Thus, in Example 4, the solution set is

$$\{(\text{Jim}, \text{Anne}), (\text{Tom}, \text{Mary}), (\text{Arthur}, \text{Jane})\}$$

and in Example 5 the solution set is

$$\{(a, 1), (a, 2), (b, 1), (b, 3), (d, 1), (d, 2), (d, 3), (e, 2)\}.$$

Notice that we have been careful to refer to “is the husband of” and “is linked by a line to” as *relationships* and not relations. This is because we want to emphasize that a relation is not defined unless we are given the set or sets concerned, just as a function is not defined just by the rule for obtaining the images, for we also need to know the domain and a codomain.

We have two possible ways of approaching the definition of a *relation*.

- (i) We may define a relation by specifying the set of ordered pairs for which the relationship concerned holds (i.e. the solution set), together with the set for which it does not hold.
- (ii) Alternatively, we may define a relation by specifying two sets (which may be equal), together with a statement in words or symbols of the relationship by which we compare elements of one set with elements of the other.

Whichever method we choose, the other defines the *same* relation. There is a third, hybrid possibility. We can specify two sets, together with a subset of their Cartesian product such that this subset is the solution set of the relation.

### Example 6

A relation from set  $A$  to set  $B$  is defined by

$$A = B = \{2, 3, 4, 8\},$$

together with a relationship:

$$a \rho b \text{ if and only if } a \text{ is a multiple of } b, \text{ i.e. there is an integer } n, n \neq 1, \text{ such that } a = n \times b \quad (a \in A, b \in B).$$



*Example 7*

A relation is defined by the solution set:

$\{(\text{Shakespeare}, \text{Hamlet}), (\text{Shakespeare}, \text{Othello}), (\text{Shaw}, \text{St. Joan}), (\text{Shaw}, \text{The Apple Cart})\}$ ,

together with the set:

$\{(\text{Shakespeare}, \text{St. Joan}), (\text{Shakespeare}, \text{The Apple Cart}), (\text{Shaw}, \text{Hamlet}), (\text{Shaw}, \text{Othello}), (\text{Rattigan}, \text{Hamlet}), (\text{Rattigan}, \text{Othello}), (\text{Rattigan}, \text{St. Joan}), (\text{Rattigan}, \text{The Apple Cart})\}$ .

*Exercise 2*

- (i) Define the relation of Example 6 in terms of sets of ordered pairs.
- (ii) Define the relation of Example 7 in terms of sets and a relationship  $\rho$  in words.

Of course, it is not always practicable to list all the elements of the sets of a relation, nor to list all the ordered pairs of the solution set. But whichever definition of a relation we adopt, we must make clear both the relationship and also the set or sets to which it applies. We now give two possible equivalent definitions of a relation.

Given two sets  $A$  and  $B$  (which may be equal), any subset\* of  $A \times B$  defines a **relation** from  $A$  to  $B$ . If  $(a, b)$  belongs to this subset, then  **$a$  is related to  $b$** . If  $(a, b)$  is an element of  $A \times B$  and does not belong to this subset, then  **$a$  is not related to  $b$** .

A **relation** is defined by two sets (which may be equal), together with a statement which is either true or false when it is used to link any member of one set with any member of the other set in a prescribed order.

(The words *prescribed order* are necessary, because the statement may be, for example, " $a$  is taller than  $b$ ".)

We may ask ourselves which of the two definitions is to be preferred. The answer is that it depends on the particular way in which our information presents itself. It is useful to have the two alternative definitions at our disposal.

You will probably have noticed that we have defined a relation so far solely in terms of the comparison of *two* elements. Strictly speaking, what we have been discussing are *binary relations*, and there are also such things as *ternary*, . . . ,  *$n$ -ary*, etc. *relations* just as there are ternary, . . . ,  *$n$ -ary*, etc. operations. By means of an  *$n$ -ary operation* (as we have seen earlier) we

\* The subset may be empty.



obtain a result from an ordered  $n$ -tuple of elements of a set. Thus a closed  $n$ -ary operation on a set  $A$  is a function:

$$\underbrace{A \times A \times \dots \times A}_{n \text{ terms}} \longrightarrow A.$$

By an  $n$ -ary relation, on the other hand, we accept or reject ordered  $n$ -tuples of elements which come from  $n$  sets (not necessarily all different).

In the remainder of this chapter we shall confine our discussions to binary relations, and so we shall, in general, drop the adjective *binary*. We shall also concentrate on relations from a set  $A$  to a set  $B$ , where  $A$  and  $B$  are the same set. In these cases, instead of writing about relations from  $A$  to  $B$ , we shall in general discuss **relations on  $A$** .

## 2.4 Types of Relations

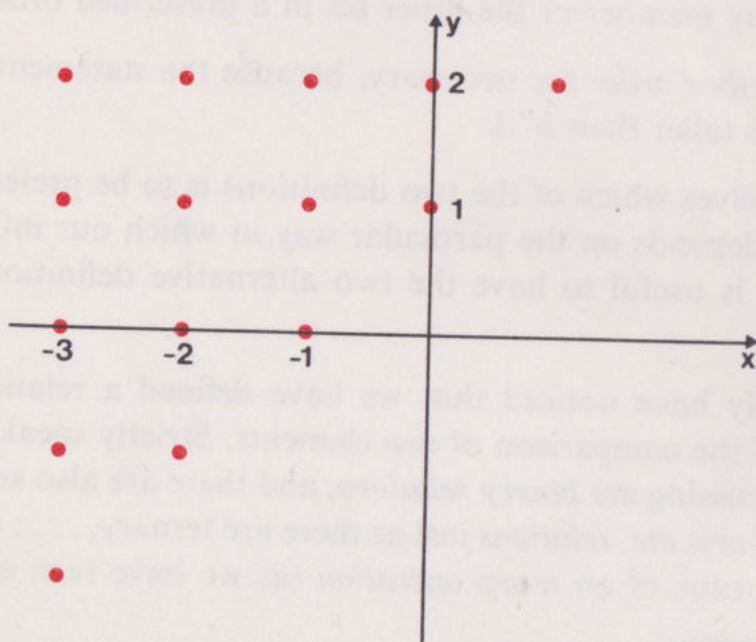
In order to single out certain properties which hold for some relations and not for others, and thus be able to classify relations according to such properties, we shall now look at a number of examples.

### Example 1

A relation on the set of integers,  $Z$ , expressed by

$$x \text{ is related to } y \text{ if } x < y \quad (x, y \in Z).$$

The solution set is a subset of  $Z \times Z$ , and corresponds to the following set of points:



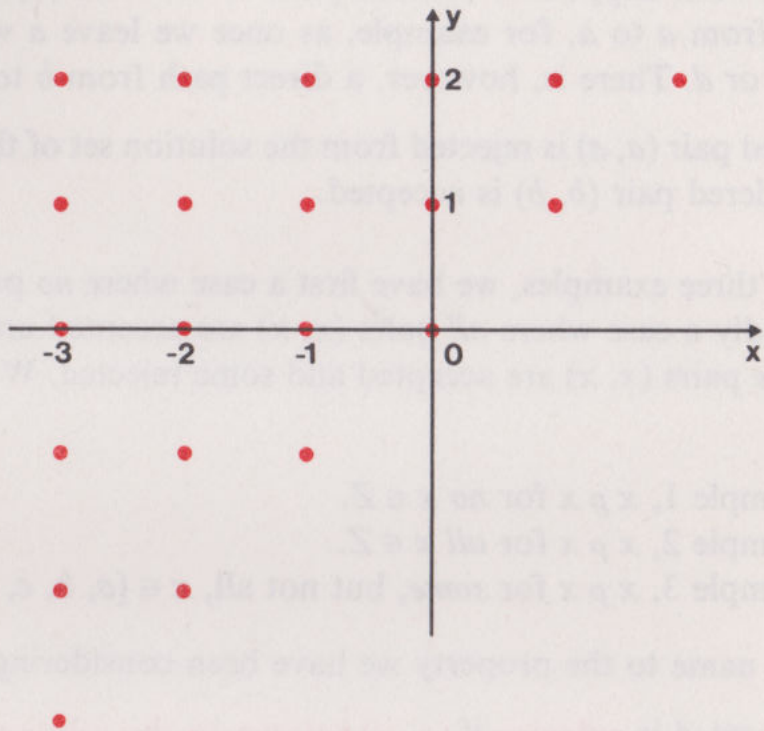


Example 2

A relation on the set of integers,  $\mathbb{Z}$ , expressed by

$$x \text{ is related to } y \text{ if } x \leq y \quad (x, y \in \mathbb{Z}).$$

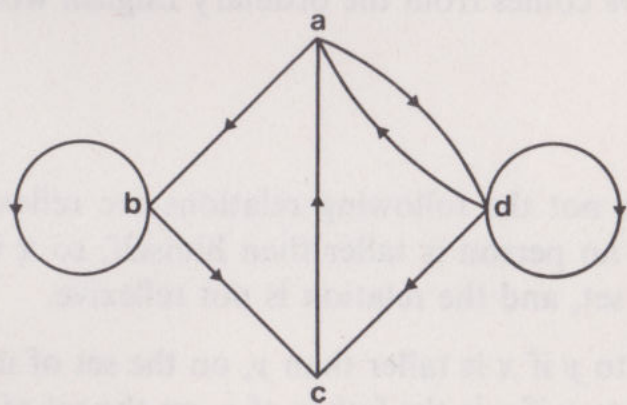
In this case we get the following representation of the solution set:



Looking at Example 1, we see that pairs such as  $(x, x)$  would be rejected from the solution set of the relation. In the case of the relation of Example 2, however, all ordered pairs of the form  $(x, x)$  would be accepted.

Example 3

Consider the diagram:



It consists of a number of points, labelled  $a$ ,  $b$ ,  $c$  and  $d$ , which are linked by paths, where the direction is indicated for each path. The paths are all, as it were, one-way streets. We can define a relation on the set of points



$\{a, b, c, d\}$  by the sentence:

“ $x$  is related to  $y$  if there is a direct path from  $x$  to  $y$  ( $x, y \in \{a, b, c, d\}$ )”.

(By a *direct* path we mean a path which does not pass through any of the other labelled points.)

Let us again see what happens to ordered pairs of the form  $(x, x)$ . There is no direct path from  $a$  to  $a$ , for example, as once we leave  $a$  we must go first to either  $b$  or  $d$ . There is, however, a direct path from  $b$  to  $b$ .

Thus the ordered pair  $(a, a)$  is rejected from the solution set of the relation, whereas the ordered pair  $(b, b)$  is accepted.

Comparing the three examples, we have first a case where *no* pair  $(x, x)$  is accepted, secondly a case where *all* pairs  $(x, x)$  are accepted and thirdly a case where *some* pairs  $(x, x)$  are accepted and some rejected. We can write this as follows:

In Example 1,  $x \rho x$  for *no*  $x \in Z$ .

In Example 2,  $x \rho x$  for *all*  $x \in Z$ .

In Example 3,  $x \rho x$  for *some*, but not all,  $x \in \{a, b, c, d\}$ .

We now give a name to the property we have been considering.

A relation on a set  $A$  is **reflexive** if  $(x, x)$  belongs to the solution set of the relation for *every*  $x \in A$ .

Alternatively, a relation  $\rho$  on a set  $A$  is **reflexive** if for all  $x \in A$

$$x \rho x$$

Thus, Example 2 is a reflexive relation, but Examples 1 and 3 are not. The term *reflexive* comes from the ordinary English word *reflex*, meaning *directed back*.

### Exercise 1

State whether or not the following relations are reflexive. For instance, in the first case, no person is taller than himself, so  $x$  is not related to  $x$  for any  $x$  in the set, and the relation is not reflexive.

- (i)  $x$  is related to  $y$  if  $x$  is taller than  $y$ , on the set of university students.
- (ii)  $x$  is related to  $y$  if  $x$  is the father of  $y$ , on the set of all people born in Great Britain.
- (iii)  $x$  is related to  $y$  if  $x$  and  $y$  are the same height measured to the nearest inch, on the set of university students.



- (iv)  $x$  is related to  $y$  if  $x$  and  $y$  are made by the same motor company, on the set of all motor cars registered in Great Britain in 1971.
- (v)  $x$  is related to  $y$  if  $x$  and  $y$  are both manufactured by B.L.M.C., on the set of all motor cars registered in Great Britain in 1971.
- (vi)  $x$  is related to  $y$  if  $x$  and  $y$  have the same derived function, on the set of all differentiable functions.

Many mathematical situations exhibit symmetry. To look for another distinguishing feature between relations we now ask:

If  $(x, y)$  is a member of the solution set, is  $(y, x)$  also a member?

Consider the following example of a relation.

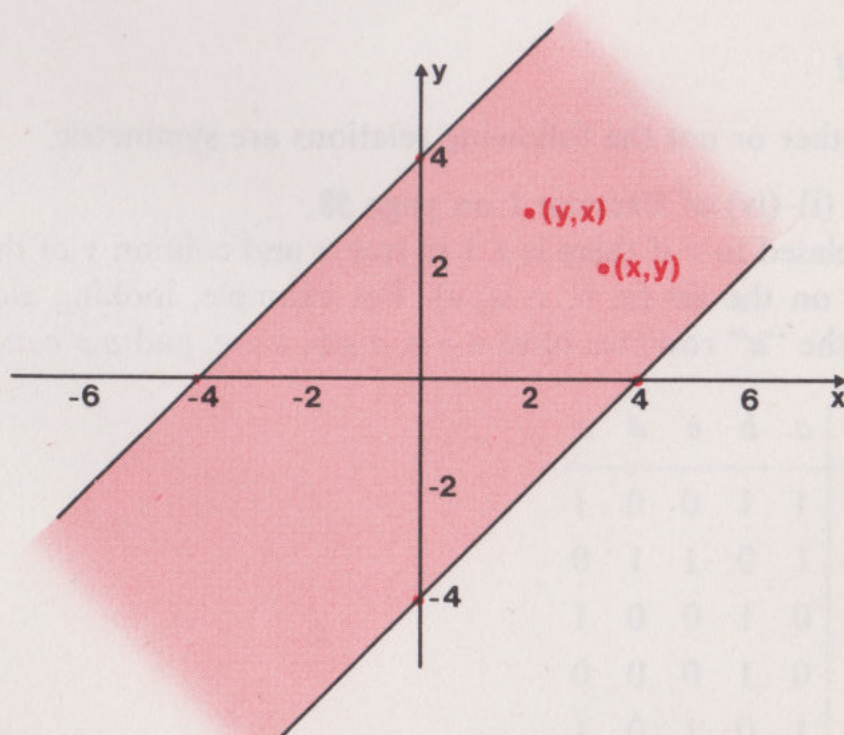
#### Example 4

We can specify a relation on the set of integers,  $Z$ , by:

$$x \text{ is related to } y \text{ if } |x - y| < 4 \quad (x, y \in Z).$$

(Remember that  $|x| = x$  if  $x \geq 0$  and  $|x| = -x$  if  $x < 0$ .)

We see that the answer to the question in this case is YES, since  $|x - y|$  is equal to  $|y - x|$ . The solution set of this relation is shown in red in the following diagram. Noticing that  $(y, x)$  is the "reflection" of  $(x, y)$  in the line with equation  $y = x$ , we can see that if  $(x, y)$  belongs to the set, so does  $(y, x)$ .





Looking back to the earlier examples, however, we find:

(i) in Example 1, where

$$x \rho y \text{ if } x < y \quad (x, y \in \mathbb{Z}),$$

that for no  $x$  such that  $x \rho y$  is  $y \rho x$ ;

(ii) in Example 2, where

$$x \rho y \text{ if } x \leq y \quad (x, y \in \mathbb{Z}),$$

that if  $x \rho y$ , then  $y \rho x$  only if  $x = y$ .

(iii) in Example 3, where

$$x \rho y \text{ if there is a direct path from } x \text{ to } y \quad (x, y \in \{a, b, c, d\}),$$

that  $a \rho d$  and  $d \rho a$ , but otherwise at most one of  $(x, y)$  and  $(y, x)$  belongs to the solution set.

We again give a name to the property we have been considering.

A relation on a set  $A$  is **symmetric** if  $(y, x)$  belongs to the solution set *whenever*  $(x, y)$  belongs to the solution set ( $x, y \in A$ ).

Alternatively, a relation on a set  $A$  is **symmetric** if

$$\text{whenever } x \rho y, \text{ then } y \rho x, \quad (x, y \in A).$$

Thus, of Examples 1–4 above, only Example 4 is a symmetric relation. The word *symmetric* refers to the property that the solution set is unchanged if the order of the elements in every pair is changed.

Exercise 2

State whether or not the following relations are symmetric.

(i) Cases (i)–(iv) of Exercise 1 on page 58.

(ii)  $x$  is related to  $y$  if there is a 1 in row  $x$  and column  $y$  of the following table, on the set  $\{a, b, c, d, e\}$ . For example, looking along the top row (the “ $a$ ” row) we have  $a \rho a, a \rho b, a \rho e$ , and  $a \not\rho c, a \not\rho d$ .

	$a$	$b$	$c$	$d$	$e$
$a$	1	1	0	0	1
$b$	1	0	1	1	0
$c$	0	1	0	0	1
$d$	0	1	0	0	0
$e$	1	0	1	0	1



We now have two properties of relations for which to look; the *reflexive* and *symmetric* properties. To consider the next property, we go back to Examples 1 and 2, namely the relations

$$x \rho y \text{ if } x < y \quad (x, y \in \mathbb{Z})$$

and

$$x \rho y \text{ if } x \leq y \quad (x, y \in \mathbb{Z}).$$

We have seen that neither of these relations is symmetric. We shall now define a property which expresses this in a more positive way.

A relation on a set  $A$  is **anti-symmetric** if *whenever*  $(x, y)$  and  $(y, x)$  both belong to the solution set, then  $x$  and  $y$  are the same element.

Alternatively, a relation on set  $A$  is **anti-symmetric** if

$$x \rho y \text{ and } y \rho x \text{ implies } x = y \quad (x, y \in A).$$

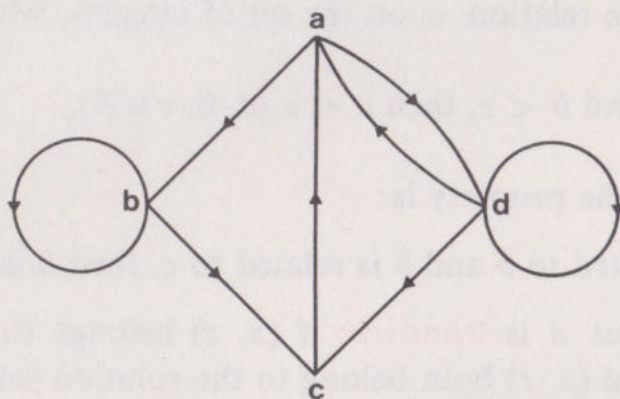
You may think at first sight that the relation of Example 1 is not anti-symmetric, but in the definition all we are saying is that pairs of the form  $(x, x)$  *may* belong to the solution set; they do not *have* to belong to it. Notice the word *whenever* in the definition; if  $(x, y)$  and  $(y, x)$  *never* both belong to the solution set, the condition is not violated.

It may seem to be implied by the terminology that it is impossible to have a relation which is both symmetric and anti-symmetric, but in fact it is possible. An example of such a relation is the relation on the set  $\{0, 1, 2, 3\}$  with solution set

$$\{(0, 0), (1, 1), (2, 2)\}.$$

This is why we use the term *anti-symmetric* rather than *not symmetric*, because it can be argued that in this special case we still have a rather trivial form of symmetry.

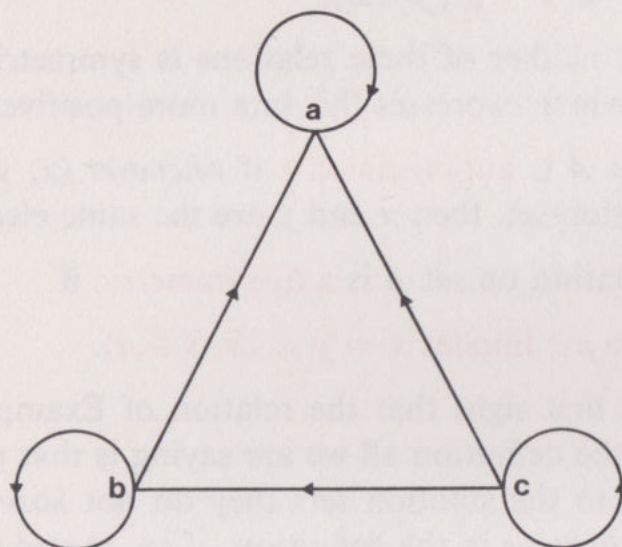
Both the relation of Example 1 and that of Example 2 are anti-symmetric, but the relation of Example 3, illustrated again in the following diagram, is neither symmetric nor anti-symmetric.





For instance,  $a \rho b$  (because there is a direct path from  $a$  to  $b$ ) but  $b \not\rho a$ , and so the relation is not symmetric. On the other hand,  $a \rho d$  and  $d \rho a$ , and so the relation is not anti-symmetric either. The following examples both illustrate anti-symmetric relations.

### Example 5



*Set*: the set of points  $\{a, b, c\}$ .

*Relationship*: there is a direct path from  $x$  to  $y$  ( $x, y \in \{a, b, c\}$ ).

### Example 6

*Set*: all the football teams in League Division 1 for the 1970–1971 season.

*Relationship*: team  $x$  is related to team  $y$  if  $x$  beats  $y$  at each meeting during the season.

(This relation is anti-symmetric in much the same way as the “ $<$  relation”—there are no elements of the form  $(x, x)$  in the solution set.)

There is one further property that we wish to consider in this section. It is illustrated by the relation  $<$  on the set of integers, where we have

$$\text{if } a < b \text{ and } b < c, \text{ then } a < c \quad (a, b, c \in \mathbb{Z}).$$

In general terms, the property is:

$$\text{if } a \text{ is related to } b \text{ and } b \text{ is related to } c, \text{ then } a \text{ is related to } c.$$

A relation on a set  $A$  is **transitive** if  $(x, z)$  belongs to the solution set whenever  $(x, y)$  and  $(y, z)$  both belong to the solution set.

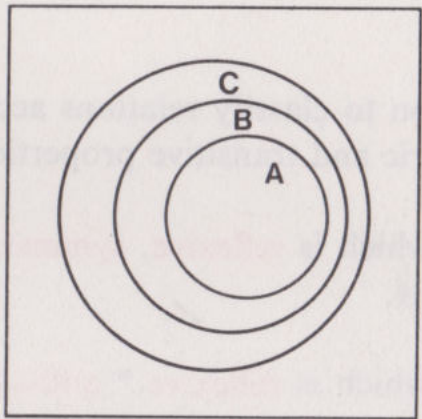


Alternatively, a relation on a set  $A$  is **transitive** if

**whenever  $x \rho y$  and  $y \rho z$ , then  $x \rho z$ .**

The term *transitive* refers to the transition from  $x$  to  $z$  via the element  $y$ .

The corresponding situation in set algebra is illustrated by the following diagram:



Since  $A \subseteq B$  and  $B \subseteq C$ , it follows that  $A \subseteq C$ .

We see that the relation defined by the relationship

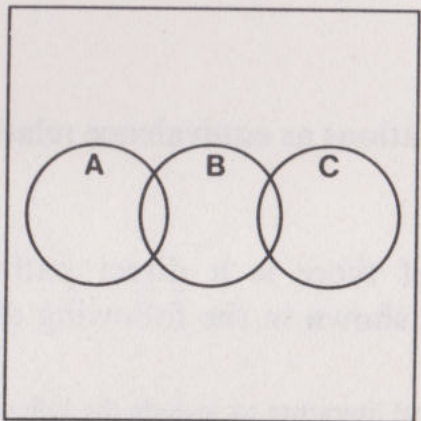
$x$  is related to  $y$  if  $x$  is a subset of  $y$ ,

on the set of sets  $\{A, B, C\}$  is transitive.

On the other hand, the relation defined by the relationship

$x$  is related to  $y$  if  $x \cap y$  is not empty,

on the set of sets  $\{A, B, C\}$ , illustrated in the following diagram, is not transitive.



In this case we have  $A \rho B$  and  $B \rho C$ , but  $A \not\rho C$ .

Another example of a transitive relation is the relation

“is parallel to”



on the set of all lines in a plane. For if a line  $L_1$  is related to a line  $L_2$ , and if  $L_2$  is related to a line  $L_3$ , then  $L_1$  is related to  $L_3$ .

### Exercise 3

State whether or not the relations in the examples in Exercise 1 on page 58 are transitive.

We are now in a position to classify relations according to the reflexive, symmetric, anti-symmetric and transitive properties.

A relation on a set  $A$  which is reflexive, symmetric and transitive is an **equivalence relation** on  $A$ .

A relation on a set  $A$  which is reflexive,\* anti-symmetric and transitive is an **order relation** on  $A$ .

None of the Examples 1–6 is an equivalence relation (you might like to see which of the required properties are lacking). The relations in (iii), (iv), and (vi) of Exercise 1 on page 58 are equivalence relations.

As far as *order relations* are concerned, the inequality relations are perhaps the most familiar examples—they enable us to arrange a set of numbers in “order of magnitude”. Of our earlier examples, 1,\* 2 and 5 are *order relations*; in each example we have a chain of elements, each element related only to all later ones. This is the familiar use of the word *order*.

### Exercise 4

Classify the following relations as equivalence relations, order relations, or neither.

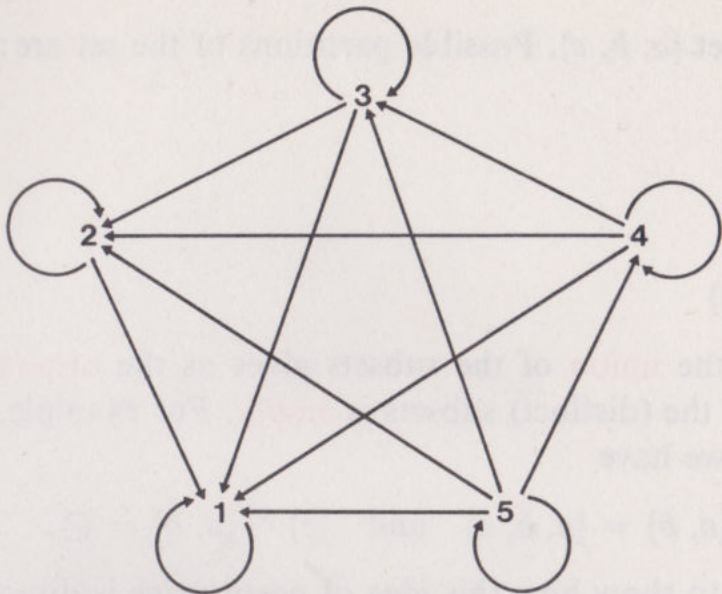
- (i)  $x$  is related to  $y$  if there is a direct path from  $x$  to  $y$  on the set  $\{1, 2, 3, 4, 5\}$ , as shown in the following diagram:

\* It is usual in the mathematical literature to include the reflexive property in the definition of an order relation: notice that this means that  $<$  on the set of integers is not strictly an order relation, although intuitively it orders the set. But if  $\rho$  is an order relation, then we can always define an associated “order” relation  $\rho_1$ , by

$$x \rho_1 y \text{ if } x \rho y \text{ and } x \neq y.$$

Then, for example, if  $\rho$  is  $\leq$ ,  $\rho_1$  is  $<$ , so we are not losing anything by the inclusion of the reflexive property in our definition.





- (ii)  $x$  is related to  $y$  if  $\sin x = \sin y$  on the set of real numbers.
- (iii)  $x$  is related to  $y$  if there is a 1 in row  $x$  and column  $y$  of the following table, on the set  $\{a, b, c, d\}$ .

	$a$	$b$	$c$	$d$
$a$	1	0	1	0
$b$	1	1	1	1
$c$	0	0	1	0
$d$	1	0	1	1

### 2.5 Equivalence Relations

In this section we shall consider equivalence relations and establish a connection with what we would usually regard as sorting. We introduce the idea of *partitioning a set*, which is the formal statement of the sorting process.

We begin by formally defining the word *partition*.

A **partition** of a set  $A$  is a **separation of the elements of  $A$  into subsets such that each element is in only one subset.**



*Example 1*

Consider the set  $\{a, b, c\}$ . Possible partitions of the set are:

- (i)  $\{a, b, c\}$
- (ii)  $\{a\}, \{b, c\}$
- (iii)  $\{b\}, \{a, c\}$
- (iv)  $\{c\}, \{a, b\}$
- (v)  $\{a\}, \{b\}, \{c\}$

In each case, the **union** of the subsets gives us the **original set**, and the **intersection** of the (distinct) subsets is **empty**. For example, in the case of partition (iv), we have

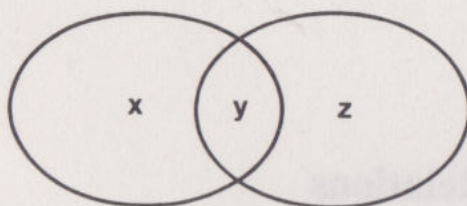
$$\{c\} \cup \{a, b\} = \{a, b, c\} \quad \text{and} \quad \{c\} \cap \{a, b\} = \emptyset.$$

We now want to show how this idea of *partitioning* is directly linked with that of an *equivalence relation*.

Given a *partition* of a set  $A$ , we can define a relation on  $A$  by:

$x \rho y$  if and only if  $x$  and  $y$  belong to the same subset of the given partition of  $A$ .

However we partition any given set  $A$ , this relation will necessarily be an equivalence relation, as can easily be checked. The transitive property is worth noting: if  $x \rho y$  and  $y \rho z$ , i.e. if  $x$  belongs to the same subset as  $y$  and  $y$  belongs to the same subset as  $z$ , then we can conclude that  $x \rho z$ , i.e. that  $x$  belongs to the same subset as  $z$ , but only because we have insisted that the subsets of a partition *do not overlap*. If the subsets were allowed to overlap, then we could have as part of our partition



in which  $x \rho y$  and  $y \rho z$  but  $x \not\rho z$ .

Let us now look at the problem the other way round and start, not with a partition of a set  $A$ , but with an *equivalence relation*  $\rho$  on  $A$ .

We can now define subsets of  $A$  by:

$x$  and  $y$  belong to the same subset of  $A$  if  $x \rho y$ .

This statement applies to all the elements of  $A$  in turn (because  $\rho$  is



reflexive and so, at least,  $x \rho x$ ), so each element must go into at least one subset of  $A$ . To show that we have a partition of  $A$ , we must show that no element of  $A$  belongs to more than one subset. Let us assume as a hypothesis the contradiction of the conjecture, that an element  $y$  belongs to two different subsets  $B$  and  $C$ , say. Suppose  $b$  is any element of  $B$  and  $c$  is any element of  $C$ .

We have  $y \rho b$  and  $y \rho c$ .

But  $\rho$  is an equivalence relation, and is therefore reflexive, symmetric and transitive. By the symmetric property we have:

$$b \rho y$$

and from  $b \rho y$  and  $y \rho c$  we have, by the transitive property,

$$b \rho c,$$

and so  $b$  and  $c$  are related and therefore  $b \in C$ .

By the symmetric property,

$$c \rho b,$$

so  $c \in B$ .

We have shown that any element of  $B$  belongs to  $C$  and any element of  $C$  belongs to  $B$ ; it follows that  $B = C$ , which contradicts the hypothesis that  $B$  and  $C$  are different.

So any equivalence relation on a set  $A$  partitions the set.

The **subsets of the partition** of a set  $A$  obtained from a given equivalence relation on  $A$  are called **equivalence classes**.

By allocating all the elements of a set uniquely to equivalence classes, we carry out in mathematical terms the process which in more general situations we have referred to as *sorting*. Thus, the mathematical model of the sorting process is the partitioning of a set by an equivalence relation into equivalence classes.

In practice, we frequently do not list all the elements of each equivalence class, and so we choose one particular element from each class as its representative. Any element belonging to a particular equivalence class may be chosen as its representative for the purpose of naming the class; for other purposes there may be external conditions which determine just how we make the choice.



*Example 2*

Consider the equivalence relation on the set of real numbers expressed in decimal form:

$x$  is related to  $y$  if  $x = y$  when both are rounded to four significant figures.

A typical equivalence class would be

$$\{x : x \in R \text{ and } 3.1415 \leq x \leq 3.1425\}.$$

Here, it is natural to select the number 3.142 as a representative of the whole class, since all the numbers in the class are rounded to this number. Knowing the equivalence relation, we can decide whether or not any number belongs to the class just by knowing this (or any other) representative.

*Exercise 1*

Give the partition of  $\{3, 4, 5, 6\}$  determined by the equivalence relation  $x \rho y$  if  $x$  and  $y$  have the same remainder on division by  $n$ , where

- (i)  $n = 2$
- (ii)  $n = 3$
- (iii)  $n = 4$ .

**The Natural Mapping**

When we partition a set under an equivalence relation, we obtain, as we have just seen, a set of equivalence classes. We call the **set of equivalence classes** the **quotient set**, and we denote it by  $A/\rho$  (which we read as “ $A$  by  $\rho$ ”), where  $\rho$  represents the equivalence relation by means of which the classes are determined.

The notation  $A/\rho$  may be a little worrying; it suggests a connection with division, as indeed does the word *quotient*.  $A/\rho$  stands for a set of subsets of  $A$ , so an element of  $A/\rho$  will be a subset of  $A$ , namely one of the equivalence classes defined by  $\rho$ . However, we know that an equivalence relation partitions a set, and so we can think of  $\rho$  as “dividing up” the set into subsets. The process of assigning an element to its equivalence class specifies a mapping from the set  $A$  to the set  $A/\rho$ . Because we have a partition of  $A$ , each element of  $A$  has an image and the sets in  $A/\rho$  are non-overlapping, so the mapping is in fact a function.



The **natural mapping**,

$$n: A \longmapsto A/\rho,$$

is the function which **maps each element of a set  $A$  to its corresponding equivalence class** under an equivalence relation  $\rho$  on  $A$ .

We may, however, find the whole situation arising the other way round. We may start with a function:

$$f: A \longmapsto B$$

in which case there is a **natural equivalence relation** on the domain  $A$  of  $f$  defined by

$$x \rho y \text{ if and only if } f(x) = f(y) \quad (x, y \in A).$$

### Example 3

Consider the function

$$f: x \longmapsto x^2 \quad (x \in R).$$

Here, each equivalence class, except that containing only the number zero, has exactly two elements, a pair of numbers of equal magnitude but of opposite sign.

### Example 4

Consider the function

$$f: x \longmapsto \sin x \quad (x \in R).$$

Here, each equivalence class contains an infinite number of elements, for example, the class of elements which map to zero under  $f$  is  $\{0, \pi, -\pi, 2\pi, -2\pi, \dots\}$ .

## Combination of Equivalence Classes

So far, we have discussed the partitioning of a set into equivalence classes, and the fact that such a partitioning leads to a particular mapping—the natural mapping. We shall now see that when a binary operation is defined on the set, it can sometimes be used to define an operation on the set of equivalence classes, a way of combining the classes themselves.

### Example 5

First consider the set of integers,  $Z$ . This set may be partitioned into those



integers which are *odd* and those which are *even* (0 is considered to be even). If we call these classes  $O$  and  $E$  respectively, then the natural mapping,  $n$ , is a many-one mapping from  $Z$  to  $\{O, E\}$ .

For example

$$n: 1 \longmapsto O$$

$$n: 156 \longmapsto E$$

Now consider an operation on  $Z$ —let us take multiplication as an example.

Compare multiplication on the set of integers with an operation  $\square$  on the set  $\{O, E\}$  defined by the following table, in which the combination  $O \square E$ , for example, is found at the intersection of the row beginning with  $O$  and the column headed by  $E$ .

$\square$	$O$	$E$
$O$	$O$	$E$
$E$	$E$	$E$

This table simply expresses the fact that *any* odd number *multiplied* by *any* odd number is odd, and so on. It represents a way of *combining the equivalence classes* corresponding to multiplication in the original set.

### Exercise 2

Construct a table which defines an operation on  $O, E$  corresponding to the operation  $+$  on the set of integers  $Z$ .

### Example 6

Consider the set of real numbers  $R$  in decimal form, and the equivalence relation on  $R$  expressed by:

$x$  is related to  $y$  if  $x = y$  when both are rounded to four significant figures.

(We saw this example before on page 68.)

For the purpose of illustration we choose the two equivalence classes

$$S_1 = \{3.124, 3.1243, 3.1238, \dots\},$$

$$S_2 = \{1.1604, 1.1597, \dots\}.$$



Consider the binary operation of addition. We have, for example,

$$3.124 + 1.1604 = 4.2844$$

which is rounded to 4.284. We also have

$$3.124 + 1.1597 = 4.2837$$

which is also rounded to 4.284. So far, it looks as though the “sum” of the classes  $S_1$  and  $S_2$  is the class containing 4.284. But is it the case that *whatever pair of numbers we choose, one from each class, their sum when rounded will always be 4.284*? If we choose 3.1243 from  $S_1$  and 1.1604 from  $S_2$  we get

$$3.1243 + 1.1604 = 4.2847,$$

which does *not* belong to the class containing 4.284. The result depends on the representatives we choose for the classes, and so *we cannot find an operation on the set of equivalence classes corresponding to  $+$  on the original set*.

Essentially what this last example is saying is that it can make a difference whether you round and then add, or whether you add and then round. And there are many more situations in mathematics which are abstractly of the same form.

### Summary

In this section we began by demonstrating that when we **partition** a set  $A$  we define an equivalence relation on  $A$ :  $x \rho y$  if  $x$  and  $y$  belong to the same subset of the partition. Conversely, an equivalence relation partitions a set into non-overlapping subsets, which we call **equivalence classes**.

We then defined the **natural mapping** as the function which maps each element to its corresponding equivalence class.

Finally, we considered a binary operation on the set  $A$  and considered the possibility of defining a corresponding binary operation on the quotient set, i.e. the set of equivalence classes.

## 2.6 Order Relations

We saw earlier that the reflexive, anti-symmetric and transitive properties are the properties which enable us to order a set; we shall discuss these properties in this section.

We remind you that a relation  $\rho$  on a set  $S$  is



*reflexive* if  $a \rho a$  for all  $a \in S$ ,

*anti-symmetric* if whenever  $a \rho b$  and  $b \rho a$  then  $a = b$

and *transitive* if whenever  $a \rho b$  and  $b \rho c$  then  $a \rho c$ .

The most familiar relation having these properties is the inequality relation  $\leq$  on a set of real numbers.

### Example 1

Consider the set of all the subsets of a given set,  $V$ , and the inclusion relation on  $V$  expressed in terms of  $\subseteq$ , i.e.  $A \subseteq B$  means that  $A$  is a subset of  $B$ . We have:

$A \subseteq A$  for all  $A \in V$  (reflexive)

$A = B$  whenever  $A \subseteq B$  and  $B \subseteq A$  (anti-symmetric)

$A \subseteq C$  whenever  $A \subseteq B$  and  $B \subseteq C$  (transitive)

for  $A, B, C \in V$ .

### Example 2

Consider the set,  $H$ , of all human beings, living or dead, and the relationship

$x \rho y$  if  $x$  and  $y$  are the same person or if  $x$  is a direct descendant of  $y$ .

We have:

$x \rho x$  for all  $x \in H$  (reflexive)

$x = y$  whenever  $x \rho y$  and  $y \rho x$  (anti-symmetric)

$x \rho z$  whenever  $x \rho y$  and  $y \rho z$  (transitive)

for  $x, y, z \in H$ .

In both these examples, the relation has the reflexive, anti-symmetric and transitive properties. A relation on a set  $A$  which is **reflexive**, **anti-symmetric** and **transitive** is a **partial ordering relation** on  $A$ .

A partial ordering relation  $\rho$  on a set  $A$  such that for all  $x, y \in A$ ,  $x \neq y$ , either  $x \rho y$  or  $y \rho x$  is called a **total ordering relation** on  $A$ .

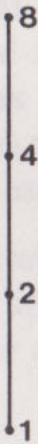
In accordance with common terminology, we shall adopt the symbol  $\leq$  in place of  $\rho$  when  $\rho$  is an order relation. Similarly, we shall use the symbol  $<$  to denote the relation  $x < y$  if  $x \leq y$  and  $x \neq y$ .

When a set has an order relation defined on it, we call it an **ordered set**.



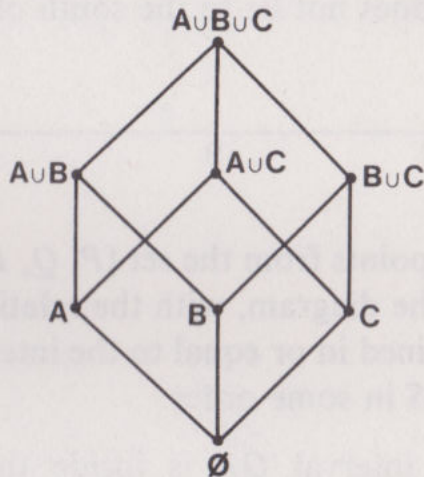
We can represent ordered sets by diagrams known as **Hasse diagrams**, some examples of which are given below:

Example 3



This could represent the set of integers  $\{1, 2, 4, 8\}$  together with  $\leq$  interpreted as “is a factor of”.

Example 4



This could represent disjoint sets,  $\emptyset, A, B, C$ , combined under the binary operation of union (as shown) to form the set  $\{\emptyset, A, B, C, A \cup B, A \cup C, B \cup C, A \cup B \cup C\}$ , with  $\leq$  interpreted as “is a subset of”.

(Remember that we consider the empty set  $\emptyset$  to be a subset of every set.)

These diagrams are an attempt to show pictorially some of the ideas we have been discussing. The ordering of the sets is illustrated by the vertical status of the elements; thus, in Example 4, the subset



$\{A \cup B \cup C, \emptyset, B, B \cup C\}$  is ordered into  $\emptyset - B - B \cup C - A \cup B \cup C$ . The subset  $\{A \cup B, B \cup C\}$  is not ordered.

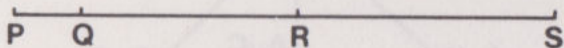
We can see from the diagram in Example 3 that the whole set is ordered—there is a link between every pair of elements. On the other hand, Example 4 is an example of *partial* ordering—there is not a link between every pair of elements. Notice that we do not have to put in *all* the direct links (i.e. Hasse diagrams are not the same as the diagrams we encountered previously, where we joined each pair of related elements). We can see that  $A$  is a subset of  $A \cup B \cup C$  because it is linked via  $A \cup B$ .

We can think, if we like, of total orderings as being a stronger property than partial ordering because it orders the whole set. Partial ordering is weaker in the sense that it orders elements within one or more subsets of the set (which need not be disjoint).

### Exercise 1

Check that the following cases give examples of ordered sets, and draw the corresponding Hasse diagrams.

- (i) The set  $\{\text{Oxford, Birmingham, London, Manchester, Exeter, Glasgow}\}$ , with the relation “does not lie to the south of”.
- (ii)



The set of pairs of points from the set  $\{P, Q, R, S\}$  of four points on a straight line as in the diagram, with the relation  $(x, y) \leq (w, z)$  if the interval  $xy$  is contained in or equal to the interval  $wz$ , where  $x, y, w, z$  represent  $P, Q, R, S$  in some order.

For example, the interval  $QR$  is inside the interval  $PS$ , so that  $(Q, R) \leq (P, S)$ ; but  $(P, R) \not\leq (Q, R)$ , because the interval  $PR$  is not included in the interval  $QR$ .

### Bounds

In various parts of Volumes 1 and 2 we considered the concept of a bound for a set of numbers.

An **upper bound** of a subset  $S$  of an ordered set  $P$  is any element  $u$  of  $P$  for which  $a \leq u$  for all elements  $a \in S$ .



We can define a *lower bound* similarly.

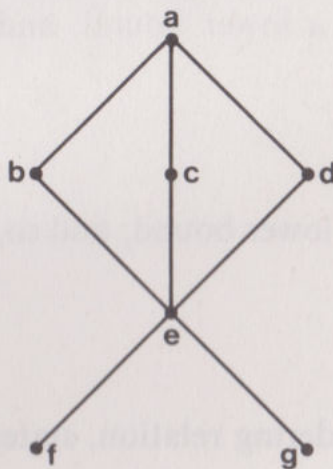
A **lower bound** of a subset  $S$  of an ordered set  $P$  is any element  $l$  of  $P$  for which  $l \leq a$  for all elements  $a \in S$ .

Notice that the upper and lower bounds must belong to the ordered set  $P$  under consideration, but *not necessarily* to the subset  $S$  in question.

Upper and lower bounds need not be unique. Thus, referring again to the diagram of Example 4, and considering the subset  $\{\emptyset, B, C\}$ , we see that both  $B \cup C$  and  $A \cup B \cup C$  are upper bounds of this subset. On the other hand, when we look for lower bounds of the same subset, we find that there is only one, namely  $\emptyset$ .

This happens to be the case here, because  $\emptyset$  is the “lowest” element of  $P$ . Were there any element “below”  $\emptyset$ , this would also be a lower bound of  $\{\emptyset, B, C\}$ . For example, the subset  $\{A \cup B, B \cup C, B\}$ , has lower bounds  $B$  and  $\emptyset$ .

*Example 5.*



Here, for the subset  $\{b, c, d, e\}$ ,  $a$  is the only upper bound, but each of  $e, f, g$  is a lower bound of the subset. If, however, we consider the set of all lower bounds of  $\{b, c, d, e\}$ , namely  $\{e, f, g\}$ , we see that one element is “higher” in the diagram than the rest, in this case  $e$ .

We call  $e$  the *greatest lower bound* of  $\{b, c, d, e\}$ .

An element  $l_g \in P$  is the **greatest lower bound** of a subset  $S$  of an ordered set  $P$ , if  $l_g$  is a lower bound of  $S$  and  $l \leq l_g$  for every lower bound  $l$  of  $S$ .

In a similar manner, we can define the *least upper bound* of a subset of an ordered set.



An element  $u_g \in P$  is the **least upper bound** of a subset  $S$  of an ordered set  $P$ , if  $u_g$  is an upper bound of  $S$  and  $u_g \leq u$  for every upper bound  $u$  of  $S$ .

Looking back to Example 5, we see, for example, that the least upper bound of  $\{b, c, e, f\}$  is  $a$ , the least upper bound of  $\{b, e, f\}$  is  $b$ , and the greatest lower bound of  $\{a, b, c, d\}$  is  $e$ .

It is important to note that the least upper bound and greatest lower bound *may not exist* for some subsets of a given partially ordered set. When they exist however, both the greatest lower bound and the least upper bound of a given subset are *unique*. (See Exercise 2.)

### Exercise 2

- (i) Consider the set  $\{1, 1 + \frac{1}{2}, 1 + \frac{1}{2} + \frac{1}{4}, 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8}, \dots\}$  considered as a subset of the reals with the relation  $\leq$  as  $\leq$ . What is the greatest lower bound and what is the least upper bound?
- (ii) By writing down the missing symbols and words, complete the following proof that, when it exists, the greatest lower bound is unique. Suppose  $u_1$  and  $u_2$  are two greatest lower bounds and  $u_1 \neq u_2$ . Then  $u_2$  is necessarily a lower bound, and so, since  $u_1$  is a *greatest* lower bound,

$$u_2 \boxed{\phantom{000}} u_1. \quad (1)$$

Also  $u_1$  is necessarily a lower bound, and so, since  $u_2$  is a *greatest* lower bound,

$$u_1 \boxed{\phantom{000}} u_2. \quad (2)$$

Since  $\leq$  is a partial ordering relation, statements (1) and (2) together imply that

$$u_1 \boxed{\phantom{000}} u_2$$

which contradicts our original assumption.

### Summary

We summarize what we have covered in this section.

First, we defined the term *ordering* and gave a useful way of representing such a relation on a set—the *Hasse diagram*. We then defined a *total ordering relation* and a *partial ordering relation*. We defined a *lower bound* and an *upper bound* of a subset of an ordered set, and found that a subset



may have more than one of each of these. It may have a *greatest lower bound* and a *least upper bound*, and if it does, each is *unique*.

There are many subjects belonging to the more recently discovered areas of mathematics, which have their beginnings in the work we have considered in this chapter.

## 2.7 Additional Exercises

### Exercise 1

- (i) If  $A$  is the set of two numbers  $\{0, 1\}$ , is the operation of addition a closed binary operation on  $A$ ?
- (ii) We were careful to say that a binary operation defines a *function*. Why is it misleading, although not inaccurate, to say that a binary operation defines a mapping?
- (iii) If  $A$  is a complete set of dominoes and we define the operation  $\circ$ , so that for example:

$$\begin{array}{|c|c|} \hline \cdot & \cdot \cdot \cdot \cdot \\ \hline \end{array} \circ \begin{array}{|c|c|} \hline \cdot \cdot \cdot \cdot & \cdot \cdot \\ \hline \end{array} = \begin{array}{|c|c|} \hline \cdot & \cdot \cdot \\ \hline \end{array}$$

(we join pieces having a number in common just as in a normal game of dominoes), is  $\circ$  a binary operation on the complete set of dominoes?

### Exercise 2

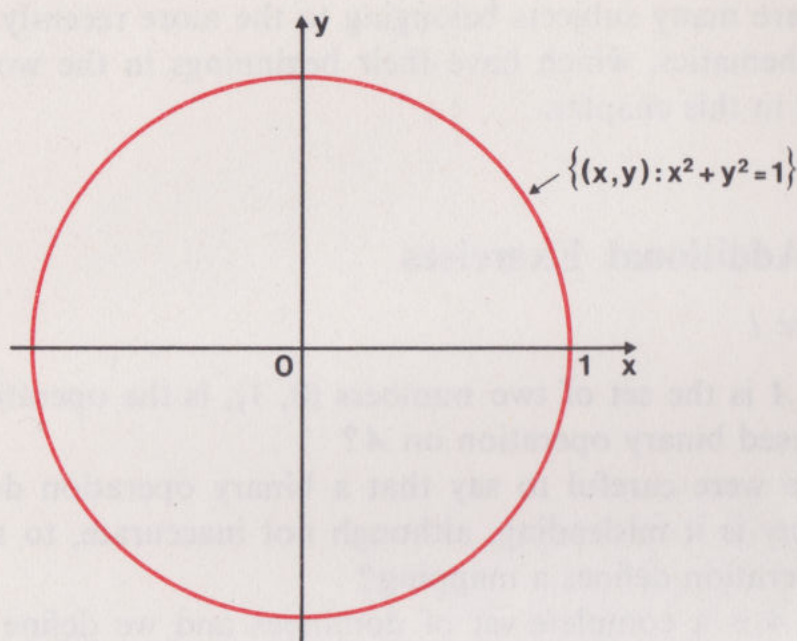
State whether or not the following relations are reflexive or symmetric.

- (i)  $x$  is related to  $y$  if there is a 1 in row  $x$  and column  $y$  of the following table, on the set  $\{a, b, c, d, e\}$ :

	$a$	$b$	$c$	$d$	$e$
$a$	1	1	1	1	1
$b$	0	1	0	0	0
$c$	1	1	1	1	1
$d$	0	0	0	1	0
$e$	1	1	1	1	1



- (ii)  $x$  is related to  $y$  if the point with co-ordinates  $(x, y)$  lies on the circle with centre the origin and radius 1, on the set  $R$ .



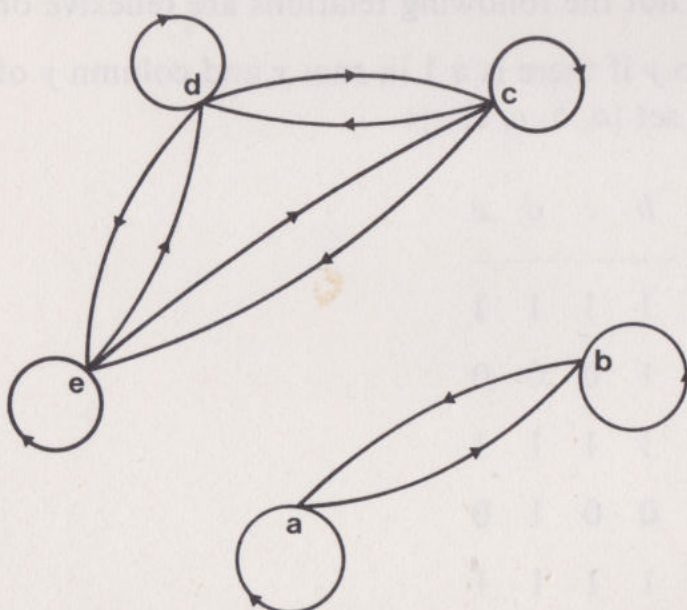
### Exercise 3

Show that the following relations are equivalence relations. Suggest how the relations can be used to sort the sets into non-overlapping subsets.

- (i) The relationship:  $x \rho y$  if  $x$  and  $y$  have the same remainder on division by 3, on the set

$\{1, 2, 3, 4, 5, 6, 7, 8\}$ .

- (ii) The relationship:  $x$  is related to  $y$  if there is a direct path from  $x$  to  $y$ , on the set  $\{a, b, c, d, e\}$ , as shown in the following diagram:

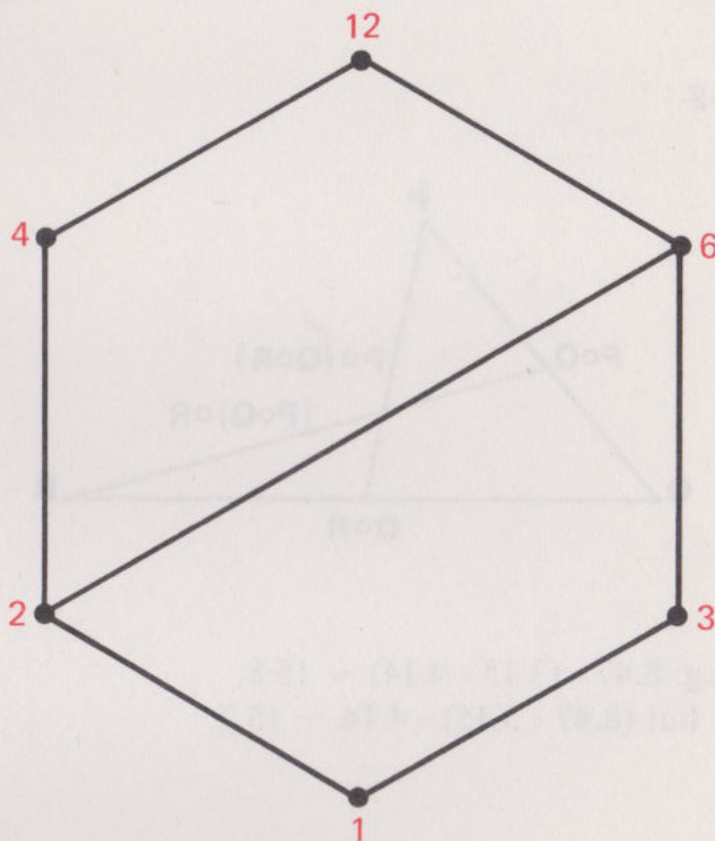




## Exercise 4

The diagram below is the Hasse diagram of the factors of 12 for the relation “is a factor of”. Show that this represents a *partial order relation*.

Give an example of a subset of at least three elements for which the relation is a *total order relation*.



## 2.8 Answers to Exercises

## Section 2.1

## Exercise 1

- (i) YES.  $a_1 + a_2$  is always a real number.
- (ii) NO. Sometimes  $a_1 - a_2$  may be a negative integer.
- (iii) NO.  $a_1 \div a_2$  need not be an integer.
- (iv) YES. The mid-point for any two points is always on the paper.
- (v) NO. The mid-point for some pairs of points may lie in the hole, which is not part of  $A$ .

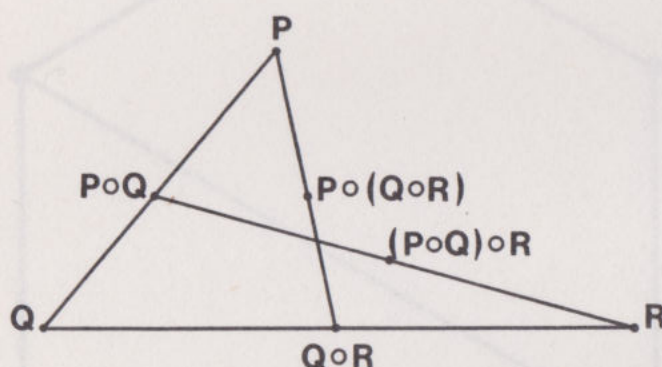
## Exercise 2

Only in (ii) and (iv) are the binary operations *not* commutative.



## Exercise 3

- (i) TRUE.
- (ii) FALSE, e.g.  $2 - (3 - 4) \neq (2 - 3) - 4$ .
- (iii) FALSE, e.g.  $(2^3)^2 = 64$ , but  $2^{(3^2)} = 512$ .
- (iv) FALSE, e.g.  $12 \div (6 \div 2) \neq (12 \div 6) \div 2$ .
- (v) TRUE.
- (vi) FALSE, e.g.:



- (vii) FALSE, e.g.  $8.47 \circ (3.15 \circ 4.14) = 15.8$ ,  
but  $(8.47 \circ 3.15) \circ 4.14 = 15.7$ .

## Exercise 4

Suppose that the operation  $\circ$  is not closed. Then, for some  $a_1, a_2 \in A$  we shall have an element  $a_1 \circ a_2$ , which will not belong to  $A$ .

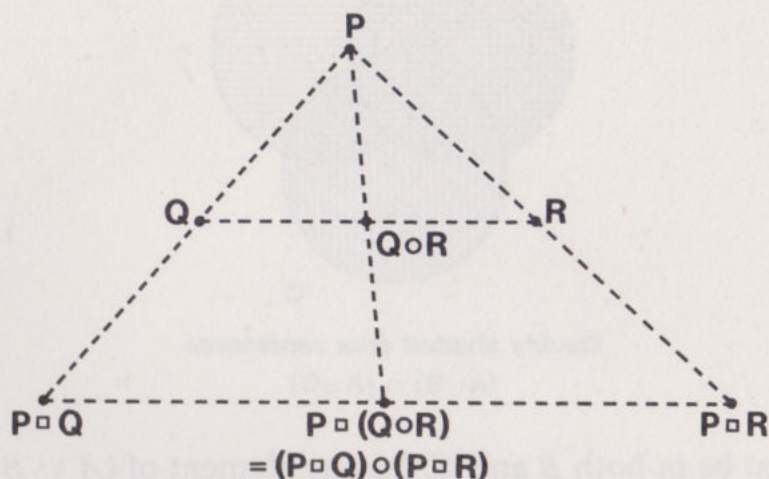
If we now try to combine this  $a_1 \circ a_2$  with  $a_3$ , we find that we are unable to do so because the operation  $\circ$  is defined *on the set*  $A$ , so we are not entitled to try to use it to combine elements not belonging to  $A$ .

## Exercise 5

- (i) FALSE, e.g.  $2 + (3 \times 4) \neq (2 + 3) \times (2 + 4)$ .
- (ii) FALSE, e.g.  $2 + (3 - 4) \neq (2 + 3) - (2 + 4)$ .
- (iii) TRUE.
- (iv) TRUE. Because multiplication is commutative, we can deduce (iv) from (iii). The order in which  $y$  and  $z$  appear has to be preserved in each case, however, because subtraction is not commutative.



(v) TRUE, e.g.



The result can be easily proved by reference to “similar triangles”.

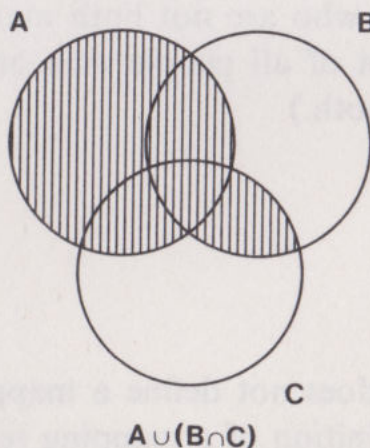
(vi) TRUE. Although, because  $\square$  is not commutative, we cannot deduce (vi) from (v), in the same way as we can deduce (iv) from (iii).

(vii) TRUE.

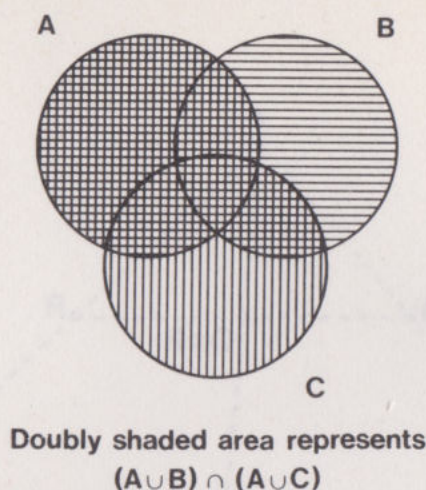
(viii) FALSE, e.g.  $8 \div (1 + 3) \neq (8 \div 1) + (8 \div 3)$ .

### Exercise 6

$A \cup (B \cap C)$  is the set of all objects that are *either* in  $A$  or in both  $B$  and  $C$ . Now any element in  $A$  is in  $A \cup B$  and in  $A \cup C$ ; and the same goes for any element that is both in  $B$  and in  $C$ ; so certainly all elements of  $A \cup (B \cap C)$  are in  $(A \cup B) \cap (A \cup C)$ . Conversely, any element of  $(A \cup B) \cap (A \cup C)$  is *both* in  $A \cup B$  and in  $A \cup C$ . Consequently, if it is







not in  $A$  it must be in both  $B$  and  $C$ . So any element of  $(A \cup B) \cap (A \cup C)$  is in  $A \cup (B \cap C)$ .

Thus,  $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ .

## Section 2.2

### Exercise 1

- (i) Let  $A$  be the set  $R$ .

The function is then  $A \times A \longrightarrow A$ , and since the set of images is  $A$ , the operation is a *closed* binary operation on  $A$ .

- (ii) Let  $A$  be the set  $R \times R$ .

The function is then  $A \longrightarrow R$ . Now the elements of  $A$  are all ordered pairs, but the elements of  $R$  are not. So  $R$  is not a subset of  $A$ , and hence the operation is a unary operation but *not* a closed unary operation.

### Exercise 2

- (i) The set of all females.
- (ii) The set of all people who do not speak English.
- (iii) The set of all males.
- (iv) The set of all people who are not both male and English-speaking.  
(Alternatively: the set of all people who are either female or non-English-speaking or both.)
- (v) The same as (iv).

## Section 2.3

### Exercise 1

The set of ordered pairs does not define a mapping because there is no “image” of Fred. The definition of a mapping requires that *each* element of the domain be mapped to an image in the codomain.



## Exercise 2

(i) Solution set:

 $\{(4, 2), (8, 2), (8, 4)\}.$ 

Other set:

 $\{(2, 2), (2, 3), (2, 4), (2, 8), (3, 2), (3, 3), (3, 4), (3, 8), (4, 3), (4, 4), (4, 8), (8, 3), (8, 8)\}.$ 

Notice that it is not sufficient simply to specify the solution set, as that set gives no indication that 3 is one of the numbers being considered.

(ii)  $A = \{\text{Shakespeare, Shaw, Rattigan}\}$  $B = \{\text{Hamlet, Othello, St. Joan, The Apple Cart}\}$  $a \rho b$  ( $a \in A, b \in B$ ) if  $a$  is the author of  $b$ .

Notice that had we not given the second set in Example 7 we would not have known that Rattigan was one of the authors belonging to set  $A$ .

## Section 2.4

## Exercise 1

(ii) Not reflexive. No person is his own father.

(iii) Reflexive.

(iv) Reflexive.

(v) Not reflexive. A Ford is not related to itself.

(vi) Reflexive.

## Exercise 2

(i) (i) No.

(ii) No.

(iii) Yes.

(iv) Yes.

(ii) Yes. The term *symmetric* is visually demonstrated in this case by the fact that the table is symmetric about the diagonal from the top left-hand corner to the bottom right-hand corner.

	a	b	c	
a	1	1	0	...
b	1	0	1	...
c	0	1	0	...
	.	.	.	
	.	.	.	
	.	.	.	



*Exercise 3*

- (i) Yes. If  $x$  is taller than  $y$  and  $y$  is taller than  $z$ , then  $x$  is taller than  $z$ .
- (ii) No.
- (iii) Yes.
- (iv) Yes.
- (v) Yes.
- (vi) Yes.

*Exercise 4*

- (i) An order relation. The relation is, in fact,  $\geq$ . We can arrange the elements in order:

5, 4, 3, 2, 1,

so that each element is related to all the later elements.

- (ii) An equivalence relation. It sorts the real numbers into subsets, every number in a particular subset having the same sine. For example, the subset  $\{0, \pi, -\pi, 2\pi, -2\pi, \dots\}$  consists of all the elements  $x$  such that  $\sin x = 0$ .
- (iii) An order relation. The elements can be arranged in order:  $b, d, a, c$ , such that each element is related to all the later elements. Each element is related to itself, so the reflexive property is satisfied. Nowhere do we get  $x \rho y$  and  $y \rho x$  where  $x \neq y$ , so the anti-symmetric property is satisfied. By checking all the possible cases, we can see that the transitive property is also satisfied. For instance,  $b$  is related to  $d$ ,  $d$  is related to  $a$ , and  $b$  is related to  $a$ .

**Section 2.5***Exercise 1*

- (i)  $\{4, 6\}, \{3, 5\}$
- (ii)  $\{3, 6\}, \{4\}, \{5\}$
- (iii)  $\{4\}, \{5\}, \{6\}, \{3\}$

*Exercise 2*

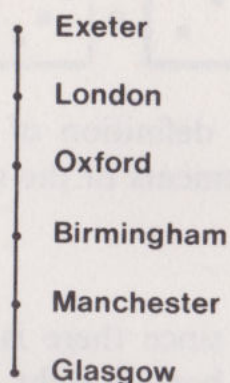
$\square$	$O$	$E$
$O$	$E$	$O$
$E$	$O$	$E$



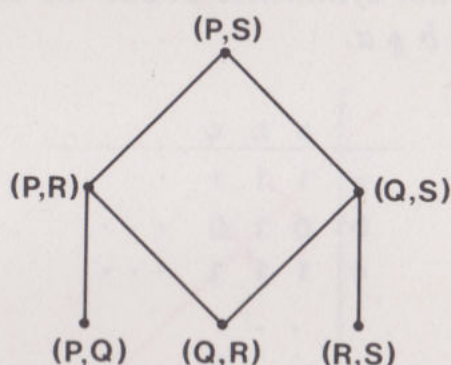
## Section 2.6

## Exercise 1

(i)



(ii)

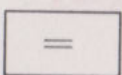
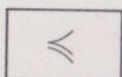
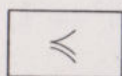


This is an example of *partial* ordering, as opposed to *total* ordering, because there is no link, for instance, between  $(P, R)$  and  $(Q, S)$ .

## Exercise 2

(i) The greatest lower bound is 1. The least upper bound is 2.

(ii)



## Section 2.7

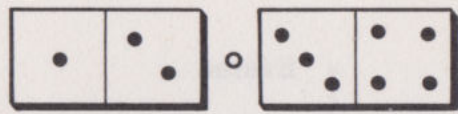
## Exercise 1

(i) NO, since  $1 + 1 = 2$  is not a member of the set  $\{0, 1\}$ .

(ii) Because our definition of a binary operation  $\circ$  states that  $a_1 \circ a_2$  is a *uniquely defined element* (see p. 38). This is the very point which distinguishes the particular mappings which are functions.



(iii) NO, since the binary operation is *not defined* for some pairs of elements, e.g.:



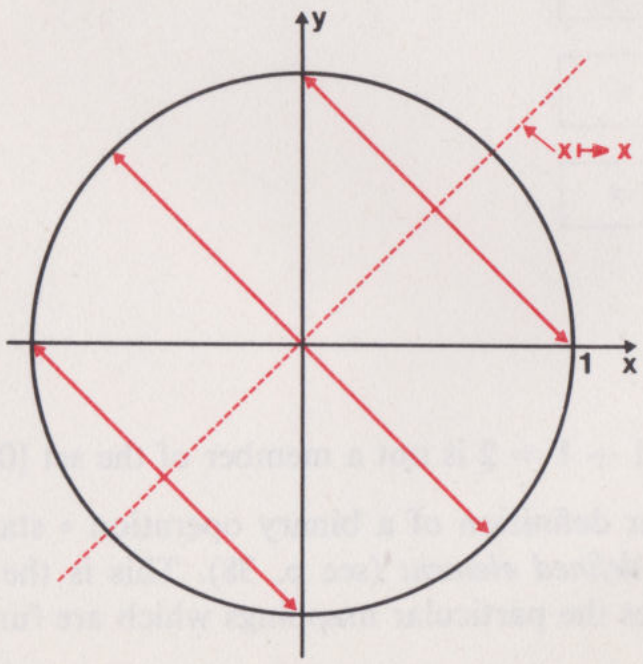
is not defined, and our definition of a binary operation on a set requires that any two elements of the set may be combined under  $\circ$ .

Exercise 2

(i) The relation is reflexive since there is a 1 in each position of the diagonal from top left to bottom right. The relation is not symmetric since the table is not symmetric about the same diagonal, and, for instance,  $a \rho b$  but  $b \not\rho a$ .

	a	b	c	
a	1	1	1	...
b	0	1	0	...
c	1	1	1	...
	...	...	...	
	...	...	...	
	...	...	...	

(ii) The relation is not reflexive since the line with equation  $y = x$  is not included in its solution set. The relation is symmetric since the circle is symmetric about the line with equation  $y = x$ . Algebraically, if  $a \rho b$ , then  $a^2 + b^2 = 1$ . This implies that  $b^2 + a^2 = 1$ , and so  $b \rho a$ .





*Exercise 3*

- (i)  $x \rho x$  for all  $x$  in the set (reflexive property).

If  $x$  leaves the same remainder as  $y$  on division by 3, then  $y$  leaves the same remainder as  $x$ . That is, if  $x \rho y$ , then  $y \rho x$  (symmetric property).

If  $x$  leaves the same remainder as  $y$ , and  $y$  leaves the same remainder as  $z$  then,  $x$  leaves the same remainder as  $z$ . That is, if  $x \rho y$  and  $y \rho z$ , then  $x \rho z$  (transitive property).

The relation is reflexive, symmetric and transitive, and it is therefore an equivalence relation.

- (ii) Each of the three properties can be checked as in (i).

We can sort the sets into non-overlapping subsets by collecting into subsets all the elements which are related to each other. For example, in (i) there will be one subset consisting of numbers which leave remainder 0, one of numbers which leave remainder 1, and one of numbers which leave remainder 2.

In (i) the set is sorted into three non-overlapping subsets of related elements,  $\{3, 6\}$ ,  $\{1, 4, 7\}$ ,  $\{2, 5, 8\}$ , and in (ii) the set is sorted into two subsets,  $\{a, b\}$ ,  $\{c, d, e\}$ , of related elements.

*Exercise 4*

The relation is one of *partial* order because, for example 6, is a factor of 12, but 6 is not a factor of 4.

Examples of subsets for which the relation is one of *total* order are  $\{1, 2, 4, 12\}$ ,  $\{1, 2, 6, 12\}$ ,  $\{1, 3, 6, 12\}$ , together with subsets of any of these subsets.



## CHAPTER 3 MORPHISMS

### 3.0 Introduction

In Chapter 2 we discussed the concept of *binary operation*. We shall now see how this concept combines with that of *function* to give us a rather more complex, though still fundamental, concept which we call a *morphism*.

We are going to ask you to spend some time in coming to grips with the morphism concept, because you will find that it is a recurring theme throughout mathematics. In this chapter we introduce the concept and see how it can illuminate the notion of a **mathematical model**. By a mathematical model we mean simply a mathematical structure which represents certain specific features of the physical world. When we look for a mathematical model, we want it to represent as faithfully as possible some physical or some other mathematical situation. Our search for the right model can be very much assisted if we understand the features that are common to all models, and know the kinds of question to which we should be seeking answers. You will find that the morphism concept highlights these common features and suggests just the right kind of questions to ask.

But mathematics provides us also with many useful opportunities of *modelling mathematics*. That is to say, having set up a mathematical problem in one particular way, we may find ourselves wanting to tackle the problem from an entirely different standpoint, because our initial statement of the problem in mathematical terms raises difficulties of understanding or of calculation. It is then that we seek an alternative approach, i.e. we look for a mathematical model of the mathematics.

We conclude the chapter with a brief look at an example of a morphism which is of some practical interest in the application of mathematics, the concept of *dimension*.

### 3.1 How Morphisms Arise

Let us suppose that we have a set with a *closed* binary operation and a *closed* unary operation. As our specific example, we shall consider the set  $R$  together with the binary operation of addition and the unary operation defined by the rule

double it
-----------

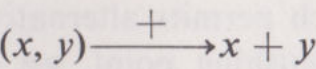


Now we know that

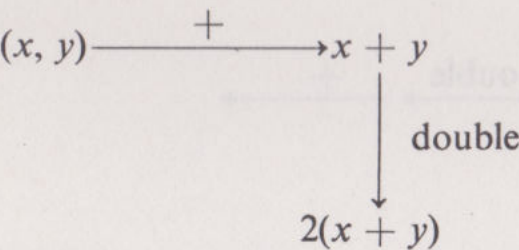
$$2(x + y) = 2x + 2y \quad (x, y \in R)$$

Equation (1)

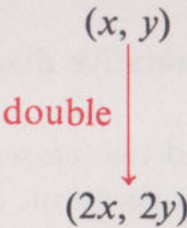
i.e. multiplication by 2 is distributive over addition. The left-hand side of our expression represents performing the binary operation first and then the unary operation, and we can represent this diagrammatically. We start with two elements of  $R$ , perform an addition, and obtain a single element of  $R$ , and we represent this by the diagram:



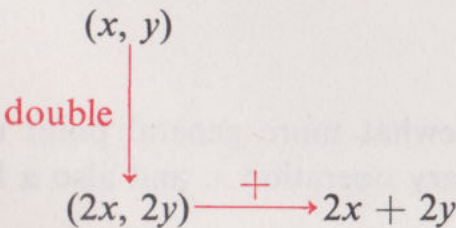
We now perform the unary operation and obtain another single element of  $R$ , and we represent the whole process by



The right-hand side of Equation (1) represents our performing the unary operation first, and this is performed upon each of the two elements of  $R$  separately, so we have



Now, performing the binary operation, we have



and because of the distributivity of multiplication over addition we arrive at the same result as before, namely  $2x + 2y = 2(x + y)$ . We can thus



put the diagrams representing the left- and right-hand sides of our expression together to give us:

$$\begin{array}{ccc}
 (x, y) & \xrightarrow{+} & x + y \\
 \text{double} \downarrow & & \downarrow \text{double} \\
 (2x, 2y) & \xrightarrow{+} & 2x + 2y = 2(x + y)
 \end{array}$$

When we have a diagram, such as this, which permits alternative paths from the same starting point to the same finishing point, we call it a **commutative diagram**. We use the word “commutative” because the order in which we perform the binary and unary operations is different in the two possible paths, and we can thus loosely think of the operations being commutative.

$$\xrightarrow{+} \xrightarrow{\text{double}} = \xrightarrow{\text{double}} \xrightarrow{+}$$

### Exercise 1

Can we draw a commutative diagram for the set  $R$  with the operations

(i)  $\times$  and square it ?

(ii)  $+$  and square it ?

If your answer is YES, draw the corresponding commutative diagram.

So far, we have taken as our starting point one set and two *closed* operations, one a binary operation and the other a unary operation. Suppose now that we drop the closure restriction on the unary operation, and, looking back to section 2.2, recall that a unary operation is merely another way of looking at a *function*

$$f: A \longrightarrow B$$

This means that we take up a somewhat more general point of view, starting with a set  $A$  with closed binary operation  $\circ$ , and also a function

$$f: A \longmapsto B = f(A)$$

Let us first see how far we can get with our diagram. We will start with



an ordered pair of elements from our set as we did before, and we shall continue to use a horizontal arrow to depict the binary operation and a vertical arrow to depict the function.

This gives:

$$\begin{array}{ccc} (a_1, a_2) & \xrightarrow{\circ} & a_1 \circ a_2 \\ \downarrow f & & \\ (f(a_1), f(a_2)) & & \end{array}$$

Since the binary operation  $\circ$  is closed, we have  $a_1 \circ a_2 \in A$ , and we can now draw a vertical arrow (depicting the function) on the right-hand side of the diagram.

$$\begin{array}{ccc} (a_1, a_2) & \xrightarrow{\circ} & a_1 \circ a_2 \\ \downarrow f & & \downarrow f \\ (f(a_1), f(a_2)) & & f(a_1 \circ a_2) \end{array}$$

Having obtained three sides of a commutative diagram, our problem now is what to do about the final side. Remembering that the missing arrow should represent the combination of  $f(a_1)$  and  $f(a_2)$  by some binary operation, we ask ourselves the question:

Is there a binary operation, say  $\square$ , on  $f(A)$  such that

$$f(a_1) \square f(a_2) = f(a_1 \circ a_2)$$

for all  $a_1, a_2 \in A$ ?

Our first reaction to this question will probably be to see if the binary operation  $\circ$ , defined on  $A$ , will do what we want. But we must remember that unless the image set  $f(A)$  is a subset of  $A$ , we are not entitled to combine elements of  $f(A)$  using  $\circ$ . In certain circumstances, as we shall see, it may be possible to *extend* our original definition of  $\circ$  so that we can combine elements of  $f(A)$  using  $\circ$  even though  $f(A)$  is not a subset of  $A$ . We now look at three particular examples.



*Example 1*

Consider the closed binary operation of addition on  $R$  and the function (unary operation):

$$f: x \longmapsto |x| \quad (x \in R)$$

the modulus function, for which

$$f(x) = x \quad \text{if } x \geq 0$$

$$f(x) = -x \quad \text{if } x < 0$$

In this case, the set of all images is the set  $R_0^+$  (the positive real numbers with zero) and, as this is a subset of  $R$ , it is clear that the operation of addition defined on  $R$  can be performed on the images in just the same way as it can be performed on the elements of  $R$ .

Let us now see how far we can get with a diagram. For  $x, y \in R$ , we have:

$$\begin{array}{ccc} (x, y) & \xrightarrow{+} & x + y \\ \text{mod} \downarrow & & \downarrow \text{mod} \\ (|x|, |y|) & & |x + y| \end{array}$$

We are allowed to try to use “+” to complete the diagram since

$$|x| + |y|$$

is defined. Unfortunately, however, it is not generally true that

$$|x| + |y| = |x + y| \quad (x, y \in R)$$

For instance, for  $x = -2$  and  $y = 2$ , we have  $|x| + |y| = 4$  and  $|x + y| = 0$ .

So, although we can combine  $|x|$  and  $|y|$  using “+”, this will not bring us to the same result as that which we obtained when we performed the addition first, and performed the modulus mapping second. Thus, we cannot in this case use + for  $\square$ .

*Example 2*

Consider the closed binary operation of addition on  $R^+$  and the function

$$f: x \longmapsto -\sqrt{x} \quad (x \in R^+)$$

We have immediately for  $x, y \in R^+$ :



$$\begin{array}{ccc}
 (x, y) & \xrightarrow{+} & x + y \\
 \downarrow -\sqrt{\phantom{x}} & & \downarrow -\sqrt{\phantom{x}} \\
 (-\sqrt{x}, -\sqrt{y}) & & -\sqrt{(x + y)}
 \end{array}$$

Now, the function is  $R^+ \mapsto R^-$  (the set of negative real numbers) and, since  $R^-$  is not a subset of  $R^+$  (the set on which we have defined our original binary operation), we cannot immediately combine elements of the image set using “+”. Addition is, however, definable on all the real numbers, and so it is readily interpreted on  $R^-$ . Once we have so extended the definition of the binary operation “+”, we can see if it can be used to complete the diagram. Unfortunately, we are again in a position exactly similar to that of the previous example, since

$$(-\sqrt{x}) + (-\sqrt{y}) \neq -\sqrt{(x + y)}$$

For example, for  $x = 9$  and  $y = 16$ , we have

$$(-\sqrt{x}) + (-\sqrt{y}) = -7 \quad \text{and} \quad -\sqrt{(x + y)} = -5$$

So, again, we cannot use + for  $\square$ .

### Example 3

Consider the closed binary operation of multiplication on  $R$  and the function

$$f: x \mapsto x^n \quad (x \in R)$$

(where  $n$  is an integer).

We have immediately for  $x, y \in R$ :

$$\begin{array}{ccc}
 (x, y) & \xrightarrow{\times} & x \times y \\
 \downarrow ( )^n & & \downarrow ( )^n \\
 (x^n, y^n) & & (x \times y)^n
 \end{array}$$

Since our original binary operation “ $\times$ ” is defined on the image set



and since, further,

$$x^n \times y^n = (x \times y)^n \quad (x, y \in R)$$

we can use  $\times$  to complete our commutative diagram and obtain:

$$\begin{array}{ccc} (x, y) & \xrightarrow{\times} & x \times y \\ \downarrow ( )^n & & \downarrow ( )^n \\ (x^n, y^n) & \xrightarrow{\times} & x^n \times y^n = (x \times y)^n \end{array}$$

### Exercise 2

For each of the given sets  $A$ , binary operations  $\circ$  on  $A$ , and functions  $f$ , comment on the use of  $\circ$  on the image set and state, when appropriate, if the commutative diagram can be completed.

Draw the completed diagram where possible.

- (i) set  $A$ :  $R$  (excluding zero)  
 binary operation  $\circ$ :  $\times$   
 function  $f$ :  $x \mapsto \frac{1}{x} \quad (x \in A)$
- (ii) set  $A$ :  $R^+$   
 binary operation  $\circ$ :  $+$   
 function  $f$ :  $x \mapsto \frac{1}{+\sqrt{x}} \quad (x \in A)$
- (iii) set  $A$ :  $Z^+$  (the set of positive integers)  
 binary operation  $\circ$ :  $\times$   
 function  $f$ :  $\left. \begin{array}{l} x \mapsto 0, \text{ when } x \text{ is even} \\ x \mapsto 1, \text{ when } x \text{ is odd} \end{array} \right\} \quad (x \in A)$
- (iv) set  $A$ :  $R$  (excluding zero)  
 binary operation  $\circ$ :  $\div$   
 function  $f$ :  $x \mapsto 1 - x \quad (x \in A)$

What we have been considering so far is a special case of the more general problem we posed originally on page 91:



Is there a binary operation  $\square$  on  $f(A)$  such that

$$f(a_1) \square f(a_2) = f(a_1 \circ a_2)$$

for all  $a_1, a_2 \in A$ ?

#### Example 4

Consider

$$f: x \mapsto \log x \quad (x \in \mathbb{R}^+)$$

and the closed binary operation of multiplication. We have:

$$\begin{array}{ccc} (x, y) & \xrightarrow{\quad \times \quad} & x \times y \\ \text{log} \downarrow & & \downarrow \text{log} \\ (\log x, \log y) & \xrightarrow{\quad \square ? \quad} & \log(x \times y) \end{array}$$

Because of the properties of the function  $x \mapsto \log x$ , ( $x \in \mathbb{R}^+$ ), we know that

$$\log(x \times y) = \log x + \log y$$

(It is this particular property that enables us to use logarithms for calculating products.) So we *can* complete our diagram using the binary operation of *addition* on the bottom line to give us:

$$\begin{array}{ccc} (x, y) & \xrightarrow{\quad \times \quad} & x \times y \\ \text{log} \downarrow & & \downarrow \text{log} \\ (\log x, \log y) & \xrightarrow{\quad + \quad} & \log x + \log y = \log(x \times y) \end{array}$$

More generally we would have:

$$\begin{array}{ccc} (a_1, a_2) & \xrightarrow{\quad \circ \quad} & a_1 \circ a_2 \\ f \downarrow & & \downarrow f \\ (f(a_1), f(a_2)) & \xrightarrow{\quad \square \quad} & f(a_1) \square f(a_2) = f(a_1 \circ a_2) \end{array}$$



Whenever we can “complete the rectangle” in this way using either the same or a different binary operation in the image set, the function  $f$  is called a **morphism** from the set  $A$  with binary operation  $\circ$ , written  $(A, \circ)$ , to the set  $f(A)$  with binary operation  $\square$ , written  $(f(A), \square)$ .\*

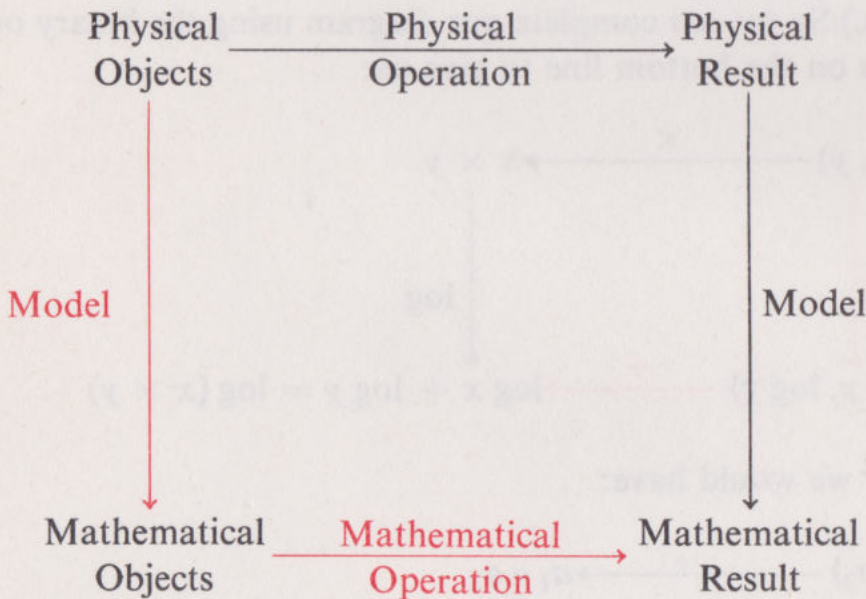
We write

$$f: (A, \circ) \longmapsto (f(A), \square)$$

In the simplest of cases  $f(A)$  is a subset of  $A$  and  $\square$  is the same as  $\circ$ , but in many of the examples which are of use in mathematical situations, such as the logarithm example, we have a *modelling* of one set with a binary operation by another set with a different binary operation.

We can, if we like, think of the function  $f: A \longmapsto f(A)$  providing us with an “image” of the binary operation defined on  $A$ , just as it provides us with images of the elements of  $A$  and of  $A$  itself. Thus, in our logarithm example, we can think of addition on  $R$  as being the image of multiplication on  $R^+$ , and we could write  $f(\times) = +$  to express this.

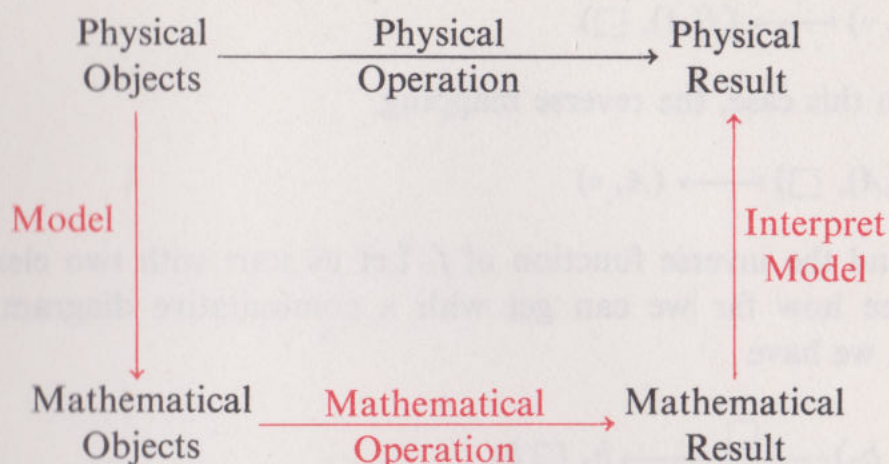
The importance of morphisms lies in their modelling property. Thus, we may have a physical situation with perhaps some mechanical or electrical operation which we model by selecting a suitable mathematical set and operation. We can then draw the diagram as:



\* Many authors define a morphism  $f$  from  $(A, \circ)$  to  $(B, \square)$ , where  $B$  is the codomain but not necessarily the image set of  $A$  under  $f$ . There is no advantage to us at this stage in this slightly more general definition.



In practice, the process used corresponds to a modified commutative diagram:



In other words, the usefulness of the model will depend very largely upon whether or not we can go from the mathematical result to the physical result. That is, the usefulness depends on our being able to reverse the right-hand arrow, and on the sort of interpretation which we can put on the mathematical results when we translate them back into the physical situation. The reversal of the arrow corresponds to the finding of the reverse mapping of  $f: A \longmapsto B$ , i.e.  $g: B \longmapsto A$ , and this is not necessarily a function.

In the case of modelling a physical situation, we have the additional problems associated with re-interpreting an idealized mathematical model, and we must always remember that our mathematical model may well represent only certain features of the physical situation, and so our mathematical result will at best be a good *approximation* to the physical result. We shall not inquire further in this chapter into the mathematical modelling of physical situations, except for the one example of dimensional analysis at the end.

(The word “morphism” is derived from the Greek word  $\mu\omicron\rho\phi\acute{\eta}$  meaning “form”; compare “metamorphosis”, meaning “changing of form”. It is used here because a morphism is a “form-preserving” or “structure-preserving” function.)

### Exercise 3

If  $f$  is a morphism  $f: (A, \circ) \longmapsto (f(A), \square)$ , is the reverse mapping  $g: (f(A), \square) \longmapsto (A, \circ)$  necessarily a morphism?

As mentioned in the solution to this exercise, in the *many-one* case, the



reverse of a morphism is not a morphism. Suppose, however, that the morphism

$$f: (A, \circ) \longmapsto (f(A), \square)$$

is one-one. In this case, the reverse mapping

$$g: (f(A), \square) \longmapsto (A, \circ)$$

is one-one and the inverse function of  $f$ . Let us start with two elements of  $B$ , and see how far we can get with a commutative diagram. For  $b_1, b_2 \in f(A)$ , we have

$$\begin{array}{ccc} (b_1, b_2) & \xrightarrow{\square} & b_1 \square b_2 \\ \downarrow g & & \downarrow g \\ (g(b_1), g(b_2)) & & g(b_1 \square b_2) \end{array}$$

Can we now complete the diagram by using the binary operation “ $\circ$ ”, i.e. does

$$g(b_1) \circ g(b_2) = g(b_1 \square b_2)?$$

So far in our analysis of this problem we have not used the obvious piece of information that  $f$  is a one-one morphism and that  $g$  is the inverse function of  $f$ . We can introduce this into the argument by supposing that

$$\left. \begin{array}{l} f(a_1) = b_1 \quad \text{i.e. that } g(b_1) = a_1 \\ f(a_2) = b_2 \quad \text{i.e. that } g(b_2) = a_2 \end{array} \right\} \begin{array}{l} \text{because} \\ f \text{ is} \\ \text{one-one} \end{array}$$

Then  $g(b_1) \circ g(b_2) = a_1 \circ a_2$ , and, using the fact that  $f$  is a morphism, we have

$$\begin{aligned} f(a_1 \circ a_2) &= f(a_1) \square f(a_2) \\ &= b_1 \square b_2 \end{aligned}$$

Again, since  $f$  is one-one and  $g$  is its inverse it follows that

$$a_1 \circ a_2 = g(b_1 \square b_2)$$

whence

$$g(b_1) \circ g(b_2) = g(b_1 \square b_2)$$



and we can therefore complete our diagram using “ $\circ$ ”. Thus, *the inverse of a one-one morphism is a morphism.*

This discussion is typical of the kind of formal argument common in modern algebra. The problem does not lie in the technical detail, but rather in mastering and ordering the facts into a coherent proof.

### 3.2 Kinds of Morphism

In the last section we encountered two particular morphisms, namely

$$\log: (R^+, \times) \longmapsto (R, +)$$

and

$$f: (Z^+, \times) \longmapsto (\{0, 1\}, \times)$$

where

$$\left. \begin{array}{ll} x \longmapsto 0 & (x \text{ even}) \\ x \longmapsto 1 & (x \text{ odd}) \end{array} \right\} (x \in Z^+).$$

(See Exercise 3.1.2 (iii).)

Notice that the former is *one-one* while the latter is *many-one*.

When the function is *one-one* the morphism is called an **isomorphism**. When the function is *many-one* the morphism is called a **homomorphism**. Thus, the function  $x \longmapsto \log x$  ( $x \in R^+$ ) is an **isomorphism** of  $(R^+, \times)$  to  $(R, +)$ . On the other hand, the function

$$\left. \begin{array}{ll} x \longmapsto 0 & (x \text{ even}) \\ x \longmapsto 1 & (x \text{ odd}) \end{array} \right\} (x \in Z^+)$$

is a **homomorphism** of  $(Z^+, \times)$  to  $(\{0, 1\}, \times)$ .

We can now restate our solution to Exercise 3.1.3 and the subsequent discussion as follows:

*The reverse of a homomorphism is not a morphism;  
the inverse of an isomorphism is an isomorphism.*

#### Exercise 1

Classify each of the following morphisms as a *homomorphism* or as an *isomorphism*.

(i)  $f: (R, \times) \longmapsto (R_0^+, \times)$

where  $f: x \longmapsto |x|$ .



$$(ii) \quad f: (A, \circ) \longmapsto (A, \circ)$$

where  $f: a \longmapsto a$ .

$$(iii) \quad f: (Z^+, +) \longmapsto (\{0, 1, 2\}, \oplus_3)$$

where  $f: x \longmapsto x_3$ , where  $x_3$  is the remainder after division of  $x$  by 3, and  $\oplus_3$  is a special kind of "addition", namely *add and then take the remainder on division by three*.

We return now to our logarithm example,

$$\log: (R^+, \times) \longmapsto (R, +)$$

which we have seen to be an *isomorphism*. Because  $f$  is *one-one*, we are able to invert the right-hand arrow of the appropriate commutative diagram and obtain each image  $g$  ( $\log x + \log y$ ) uniquely. We call  $g$  the anti-logarithm, and we thus have as our diagram:

$$\begin{array}{ccc} (x, y) & \xrightarrow{\quad \times \quad} & x \times y \\ \text{log} \downarrow & & \uparrow \text{anti-log} \\ (\log x, \log y) & \xrightarrow{\quad + \quad} & \log x + \log y \end{array}$$

We are able in this case to go from  $(x, y)$  to  $x \times y$  without carrying out the operation of multiplication explicitly, by following the process represented by the other three sides of our rectangle, each one giving us a unique result, i.e. we perform

$$\xrightarrow{\quad f \quad} \xrightarrow{\quad + \quad} \xrightarrow{\quad g \quad}$$

and it is this uniqueness of our result that enables us to multiply numbers using the addition of logarithms.

In practice, we can do a similar rearrangement of the commutative diagram when the morphism  $f$  is a homomorphism; only it is not just as simple as reversing the right-hand arrow, because, since  $f$  is many-one, when we reverse the arrow (i.e. use the reverse mapping), we are likely to get more than one element in the image at the top right. This difficulty is usually resolved by external considerations. For example, suppose we use logarithms recording only the decimal part (much in the way one uses a slide rule), e.g. for the image of 20 we record  $\log 2$ . Then instead



of an isomorphism we have a homomorphism, but we can still proceed this far:

$$\begin{array}{ccc}
 (28, 39) & \xrightarrow{\times} & \text{approx. } 30 \times 40 = 1200 \\
 \text{log} \downarrow & & \\
 (\log 2.8, \log 3.9) & \xrightarrow{+} & \log 2.8 + \log 3.9
 \end{array}$$

and the anti-log of  $(\log 2.8 + \log 3.9)$  will give the right answer except for the position of the decimal point, and this is settled by our approximate value of the answer.

In our logarithm example, we were in a situation where we, as it were “happened to know” that addition was the binary operation defined on the image set which would enable us to complete our commutative diagram. We were not, however, in a situation where we had any room to manoeuvre; *no other operation on the image set would do*. Thus, once the function together with its domain and the binary operation on that domain are prescribed, the binary operation on the image set is also determined, and for this reason we call it the **induced binary operation**. It is possible to develop a useful condition for the existence of the induced binary operation,\* but we shall not do so here.

### 3.3 Units and Dimensions

We are concluding this chapter with a brief introduction to the very important practical topic of **dimensions**. We have included this topic here because it provides us with a particularly good illustration of the morphism concept. Because dimensions are directly related to the concept of measurement, we begin with a short discussion of the idea of a **unit of measurement**.

$$A \text{-----} B$$

We cannot express the length of the line  $AB$  in terms of a number alone; for example, to say that its length is 4.5 is quite meaningless. You would

\* See M100 *Mathematics Foundation Course Unit 3, Operations and Morphisms*, The Open University Press, 1971.



not know whether we meant 4.5 inches, 4.5 centimetres, or 4.5 times the length of some previously given line, and even if we said “4.5 cm” (say), this would be meaningful only if we were already agreed on what we mean by a centimetre.

Before we can express our measurement meaningfully, then, we need to choose a *unit of length*. Our choice can be quite arbitrary, as long as it is properly defined. Two units which are in common use are the yard and the metre, and these have to be defined in terms of a *standard length*. Originally, a standard length was the distance between two parallel lines engraved on a particular bar of metal (bronze in the case of the yard, and platinum-iridium in the case of the metre). Since 1960, however, the length of the metre has been defined in terms of the wavelength of orange light emitted under specific conditions by a krypton atom of mass 86; and since 1963, the yard has been defined in terms of the standard metre.

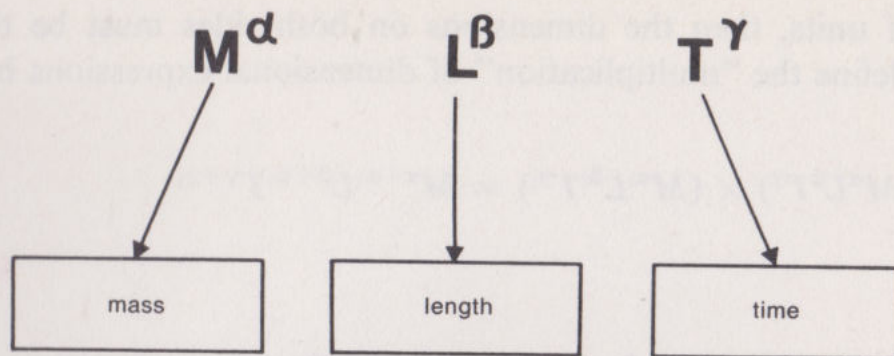
Once we have our standard length, then we can repeat it along a straight line, and so obtain lengths of 2, 3, . . . ,  $n$  units, and we can also obtain intermediate lengths of  $p/q$  units where  $p, q \in \mathbb{Z}^+$ .

We can obtain units for measuring other physical quantities in a similar manner. For example, we can define our unit of mass as the *kilogram* and our unit of time as the *second*. These two units are defined *without reference to our previously defined unit of length*, that is to say, the units of length, mass and time are all mutually *independent*.

It would be possible to choose independent units for all measurable physical quantities, but it is much more convenient, where we can do so, to choose units which are derived from the basic units of mass, length and time. For velocity, for example, we may choose as our unit the “metre-per-second”, which is a unit derived from the metre as a unit of length and the second as a unit of time. By choosing our unit of velocity in this way, we ensure that an object travelling with unit velocity, travels a unit of length in a unit of time. If we now change either of the basic units, the derived unit will be changed also.

Suppose now that we have some physical quantity,  $\phi$  (say);  $\phi$  may be velocity, or acceleration, or force, or work, etc. Let us suppose further that our derived unit of  $\phi$  is to depend solely upon our assumed units of length, mass and time. (This is an assumption because, for instance, these units are inadequate for some quantities in, say, electricity.) We now define the **dimensions** of  $\phi$  to be the symbols





where  $\alpha$ ,  $\beta$ ,  $\gamma$  will have numerical values which are appropriate for  $\phi$  in a consistent set of units. (Alternatively, we can define the **dimensions** of  $\phi$  to be the ordered triple  $(\alpha, \beta, \gamma)$ .)

### Example 1

Let  $\phi$  be velocity.

Our derived unit requires that unit length is travelled in unit time, as we have already seen.

If we increase the unit of length and keep the unit of time constant, then the new derived unit of velocity must be *increased* by the same factor. If we increase the unit of time and keep the unit of length constant, then the unit of velocity must be *decreased* by the same factor.

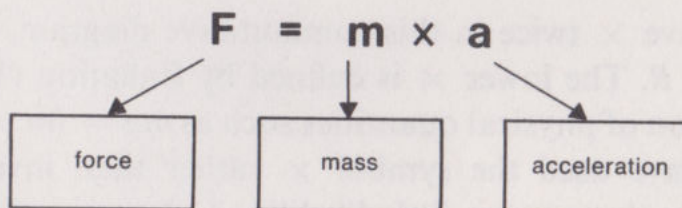
Our unit of  $\phi$  is therefore in *direct ratio* to our unit of length and in *inverse ratio* to our unit of time. Hence,  $\beta = 1$  and  $\gamma = -1$  for  $\phi$ , and, since mass has no influence,  $\alpha = 0$ . The dimensions of  $\phi$  are thus

$$M^0 \quad L^1 \quad T^{-1} \quad (\text{or } (0, 1, -1))$$

or just  $LT^{-1}$ .

### Exercise 1

- (i) What are the dimensions of *acceleration*?
- (ii) From the equation



deduce the dimensions of *force*.

If an equation relating various physical quantities is to be true in every



system of units, then the dimensions on both sides must be the same. We can define the “multiplication” of dimensional expressions by

$$(M^\alpha L^\beta T^\gamma) \times (M^{\alpha_1} L^{\beta_1} T^{\gamma_1}) = M^{\alpha+\alpha_1} L^{\beta+\beta_1} T^{\gamma+\gamma_1}$$

or

Equation (1)

$$(\alpha, \beta, \gamma) \times (\alpha_1, \beta_1, \gamma_1) = (\alpha + \alpha_1, \beta + \beta_1, \gamma + \gamma_1)$$

We thus have a simple preliminary method of checking whether or not a given equation stands a chance of being correct, and also of determining the dimensions of constants of proportion which are often encountered in relations between physical quantities. This subject of dimensional analysis is a study on its own which we do not intend to discuss more than is necessary for our present purpose.

We are able to use dimensional analysis because of the *morphism* from physical quantities to their respective dimensions which preserves “multiplication”.

We can represent this morphism by

$$\begin{array}{ccc} (\phi, \psi) & \xrightarrow{\times} & \phi \times \psi \\ \text{dim} \downarrow & & \downarrow \text{dim} \\ (\dim(\phi), \dim(\psi)) & \xrightarrow{\times} & \dim(\phi \times \psi) = \dim(\phi) \times \dim(\psi) \end{array}$$

where  $\phi, \psi \in P$ , the set of physical quantities, and “dim” is the function which maps a given physical quantity to its dimensions.

Although we have  $\times$  twice in this commutative diagram, neither is quite the usual  $\times$  on  $R$ . The lower  $\times$  is defined by Equation (1), the upper  $\times$  is the combination of physical quantities such as  $ma = (m \times a)$  in  $F = ma$ . However, we have used the symbol  $\times$  rather than invent two special symbols, because there seems little likelihood of any confusion arising.

Before asking you to attempt the final exercise, we give a list of dimensions of some mechanical quantities.



$P$	$\dim(P)$
<i>Quantity</i>	<i>Dimensions</i>
Velocity	$LT^{-1}$
Acceleration	$LT^{-2}$
Force	$MLT^{-2}$
Work (Energy)	$ML^2T^{-2}$
Angle	Dimensionless
Angular Velocity	$T^{-1}$
Angular Acceleration	$T^{-2}$
Moment of Inertia	$ML^2$

(Here “dimensionless” means that  $\dim(\text{angle}) = M^0L^0T^0$ , that is the unit of angle does not depend on mass, length or time.)

### Exercise 2

- (i) Given the equation of gravitation

$$F = G \frac{m_1 m_2}{d^2}$$

where

$F$  is force,

$m_1, m_2$  are masses,

$d$  is the distance apart of the masses,

$G$  is a gravitational constant,

find the dimensions of  $G$ .

- (ii) How should we define our unit of force so as to make  $G$ , in (i) above, identically equal to unity in all consistent systems of units?
- (iii) Why is dimensional analysis alone unable to give us the numerical value of any particular constant such as  $G$ ?

## 3.4 Conclusion

Although the morphism concept is basically a very simple idea, it does seem to cause some initial difficulty. We have, therefore, not set any additional exercises; we shall allow the idea to develop gradually over the succeeding chapters.



If you have worked through the previous volumes you will have come across a number of morphisms; for example, the following can all be regarded as statements about morphisms:

- (i)  $D(f + g) = Df + Dg$  (Volume 1, Chapter 8)
- (ii)  $\int_a^b (f + g) = \int_a^b f + \int_a^b g$  (Volume 1, Chapter 7)
- (iii)  $\lim (\underline{u} + \underline{v}) = \lim \underline{u} + \lim \underline{v}$   
 $\lim (\underline{u} \times \underline{v}) = \lim \underline{u} \times \lim \underline{v}$  (Volume 1, Chapter 6)

We shall meet quite a few more examples in the remainder of this volume. The fact that we have so many examples makes it worth while studying morphisms explicitly.

## 3.5 Answers to Exercises

### Section 3.1

#### Exercise 1

- (i) YES,  $(x \times y)^2 = x^2 \times y^2$  ( $x, y \in R$ )
- (ii) NO,  $(x + y)^2 \neq x^2 + y^2$  ( $x, y \in R$ )

For (i) we have the commutative diagram

$$\begin{array}{ccc}
 (x, y) & \xrightarrow{\times} & x \times y \\
 \text{square} \downarrow & & \downarrow \text{square} \\
 (x^2, y^2) & \xrightarrow{\times} & (x \times y)^2 = x^2 \times y^2
 \end{array}$$

#### Exercise 2

- (i) The elements of the image set may be combined using “ $\times$ ”.  
 YES, a commutative diagram can then be formed.

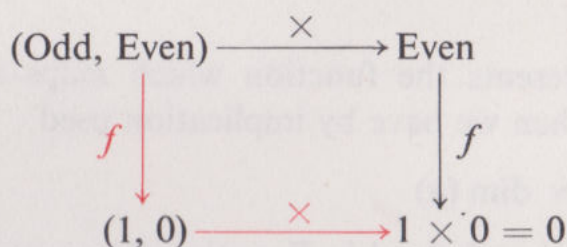
$$\begin{array}{ccc}
 (x, y) & \xrightarrow{\times} & x \times y \\
 f \downarrow & & \downarrow f \\
 \left(\frac{1}{x}, \frac{1}{y}\right) & \xrightarrow{\times} & \frac{1}{x} \times \frac{1}{y} = \frac{1}{x \times y}
 \end{array}$$



- (ii) The elements of the image set may be combined using “+”.  
NO, a commutative diagram cannot be formed, e.g.

$$\frac{1}{+\sqrt{4}} + \frac{1}{+\sqrt{9}} \neq \frac{1}{+\sqrt{(4+9)}}$$

- (iii) The binary operation “ $\times$ ” is defined on  $Z^+$  and zero is not a member of this set. The operation must therefore first be extended to the set of images  $\{0, 1\}$ . We illustrate the diagram for one of the cases:



- (iv) The function maps  $R$  (excluding zero) to  $R$  (excluding 1), and division is not defined on any subset of  $R$  which includes zero. We thus cannot use  $\div$  to combine elements of the image set.

### Exercise 3

NO.

If the function  $f$  is *many-one*, then the reverse mapping  $g$  is one-many and *not a function*. Since a morphism, by definition, is a function (with special properties),  $g$  cannot be a morphism.

## Section 3.2

### Exercise 1

- (i) Homomorphism; e.g.

$$-2 \mapsto 2$$

and

$$2 \mapsto 2$$

i.e. the function is many-one.

- (ii) Isomorphism. If  $a \neq b$ , then  $f(a) \neq f(b)$ , i.e. the function is one-one.  
(iii) Homomorphism; e.g.

$$1 \mapsto 1$$

$$4 \mapsto 1$$



$7 \mapsto 1$ , etc.

i.e. the function is many-one.

### Section 3.3

#### Exercise 1

- (i)  $LT^{-2}$  (or  $(0, 1, -2)$ )
- (ii)  $MLT^{-2}$  (or  $(1, 1, -2)$ )

(Notice that, if “dim” represents the function which maps a physical quantity to its dimensions, then we have by implication used

$$\dim(F) = \dim(m) \times \dim(a)$$

where  $\times$  is the binary operation defined in Equation (1) on page 104.)

#### Exercise 2

- (i)  $M^{-1}L^3T^{-2}$ . Since the dimensions on both sides must be the same, we have

$$MLT^{-2} = M^\alpha L^\beta T^\gamma \frac{M^2}{L^2}$$

where the dimension of  $G$  is represented by  $M^\alpha L^\beta T^\gamma$ . This gives  $\alpha = -1$ ,  $\beta = 3$ ,  $\gamma = -2$  by equating powers of  $M$ ,  $L$ ,  $T$  separately.

- (ii) Define it as the magnitude of the forces exerted on each other by two unit masses unit distance part.
- (iii) Because the function  $\dim: P \mapsto \dim(P)$  is a homomorphism. Thus any measurement of length, e.g.

3 metres,  
6 metres,  
8 feet,  
etc.

maps simply to  $L$ , and the information contained in the numerical value is lost.



## CHAPTER 4 GEOMETRIC VECTORS

### 4.0 Introduction

In the first chapter of this book we introduced the basic concept of a set. When we introduce an operation which allows us to combine elements of a set (such as addition on the set of real numbers), then we have a mathematical structure and things become more interesting. With two operations (such as addition and multiplication) the structure becomes even more interesting, and more so when we introduce mappings from the original set to another set. At each stage we make the structure more intricate, and possibly more intriguing. At some stage it may be possible to prove results which are not obvious, and indeed, results which are completely unexpected. These results may also have useful applications.

Our intention in this chapter and the next is to build a particular mathematical structure called a *vector space*. We begin with a set of elements which we call *arrows*; we then use arrows to define the set of *geometric vectors* and we discuss ways of combining geometric vectors.

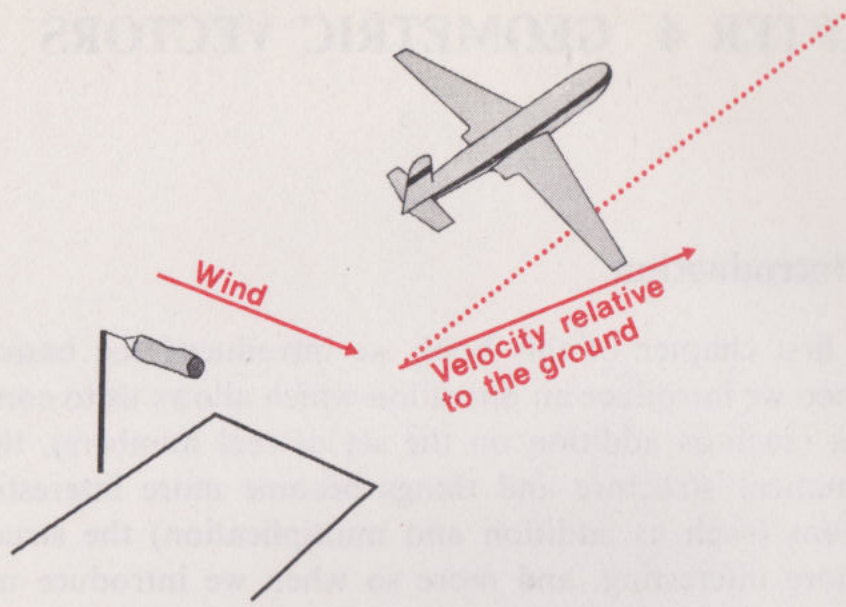
This chapter is devoted to geometric vectors for three reasons: you may have met this example of a vector space before; it is a geometric example of a vector space and many students find a pictorial approach to mathematics helpful; this geometric example is often an aid to intuition when dealing with vector spaces which are not geometric in character.

If you have met vectors before, then we advise you to notice particularly our terminology. We call the vectors which commonly arise in applied mathematics (sometimes defined as “directed line segments”) *geometric vectors*, and we use the word *vector* for an element of what we call a *vector space*.

### 4.1 Geometric Vectors

We all know that the speed of an aeroplane relative to the ground is affected by the speed of the wind. With the wind behind it, it flies faster; in a cross-wind the pilot must aim the aircraft slightly into the wind in order to reach his destination. The applied mathematician often has to deal with a situation like this in which he needs to make a mathematical model of the physical situation.

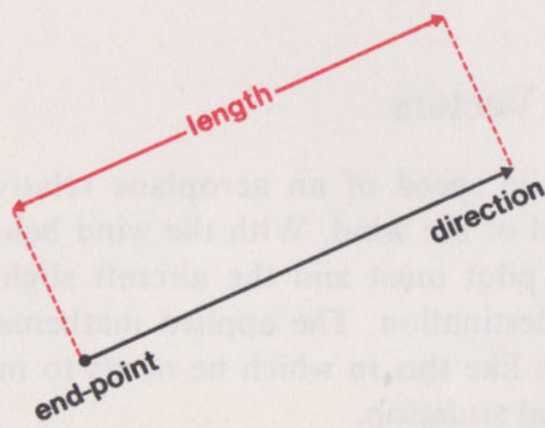




He assumes that the aircraft is being moved bodily along with the wind, just as a puff of smoke might be, and that he can represent this motion by an **arrow**, the length representing the wind speed and the direction representing the wind direction. (Incidentally he also assumes that the aircraft is compressed to a point, since its shape and size is not important for the problem under investigation.)

The speed of the aircraft in the moving air stream (the air speed) can be represented by another arrow. We shall see that the way we can combine these arrows mathematically, by considering them as representatives of geometric vectors, enables the applied mathematician to determine the direction in which the aircraft should point.

As far as we are concerned, an **arrow** has length, direction, and position. The position can be specified by one end-point, the blunt end, say.

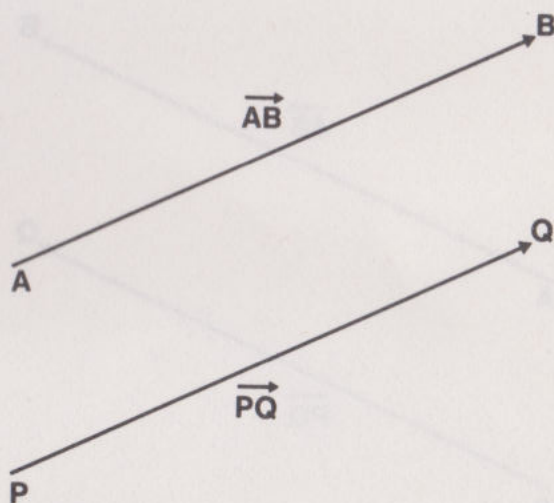




Sometimes we restrict all the arrows to lie in a plane (in other words, they can all be drawn on a flat piece of paper). On other occasions they can be in three-dimensional space but, unless it is clear from the context, we shall have to declare which case we mean at the outset of a discussion. For the moment everything we say applies to both cases.

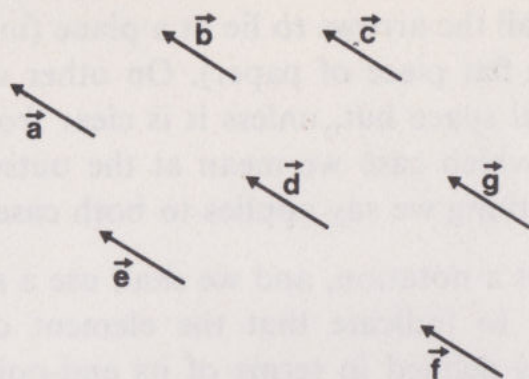
Our first requirement is a notation, and we shall use a small arrow over a letter, for example  $\vec{a}$ , to indicate that the element denotes an **arrow**. Sometimes an arrow is defined in terms of its end-points, say  $A$  and  $B$ , in which case we can write  $\overrightarrow{AB}$  for the arrow with its blunt end at  $A$  and sharp end at  $B$ .

We define two arrows to be **equal** if they have the **same length**, **direction** and **position**. It is important to notice that the two arrows  $\overrightarrow{AB}$  and  $\overrightarrow{PQ}$  shown below are distinct, even though they have the same length and direction. (They may, for example, represent the velocities associated with distinct particles at the points  $A$  and  $P$ .)

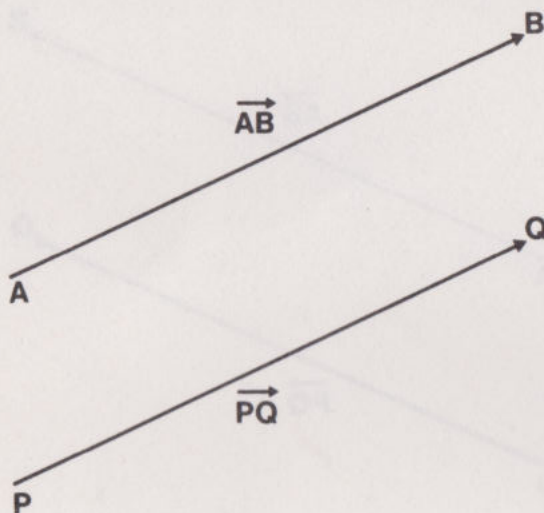


The set of *all* arrows with the same length and direction as some given arrow, but with different positions, is of interest to us because sometimes we wish to impart information in bulk, as it were, rather than for each individual arrow. (For example, we may wish to represent “a Force 6 easterly wind”.) We call such a set a **geometric vector**. Below, we have illustrated some of the arrows belonging to the same geometric vector. Remember that a geometric vector is the set of *all* arrows with the same length and direction as some given arrow, and although each of the arrows in the diagram is distinct, any one of them will determine the same geometric vector.





We need a notation for geometric vectors. Since any arrow, say  $\overrightarrow{AB}$ , is sufficient to specify the geometric vector to which it belongs, we choose to denote that geometric vector by  $\underline{AB}$ . Since geometric vectors are sets of arrows, the equality of two geometric vectors is defined: every arrow belonging to one set must also belong to the other. The two arrows  $\overrightarrow{AB}$  and  $\overrightarrow{PQ}$  belong to the same geometric vector, which we can denote by  $\underline{AB}$  or  $\underline{PQ}$ ; these geometric vectors are equal, so we write  $\underline{AB} = \underline{PQ}$ .



We use a similar notation,  $q$ , for the geometric vector of which  $\vec{a}$  is a member; that is,  $\vec{a} \in q$ .

It is impossible to draw a picture which includes *every* arrow belonging to a geometric vector, so when we need a pictorial representation of a geometric vector we draw just one of its arrows, with the understanding that it is a representative from the set.

The set of all geometric vectors is the starting point for the construction of our algebra of vectors. Before we can usefully combine geometric vectors, we need to ensure that the particular arrows chosen as representatives do not affect the result.



The relation:

has the same length and direction as

is an equivalence relation on the set of all arrows.

If we abbreviate

$\vec{a}$  has the same length and direction as  $\vec{b}$

to

$\vec{a} \rho \vec{b}$

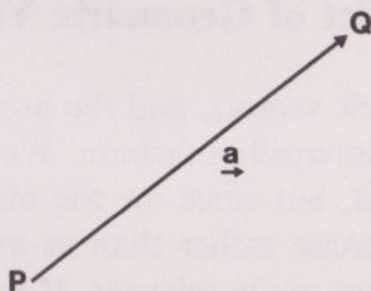
then we have:

- (i)  $\vec{a} \rho \vec{a}$  for all  $\vec{a}$ ;
- (ii)  $\vec{a} \rho \vec{b}$  implies  $\vec{b} \rho \vec{a}$ ;
- (iii)  $\vec{a} \rho \vec{b}$  and  $\vec{b} \rho \vec{c}$  implies  $\vec{a} \rho \vec{c}$ ;

that is, the three requirements of an equivalence relation are satisfied. This means that  $q$  is the equivalence class which contains the element  $\vec{a}$ . So we see that we can regard geometric vectors as equivalence classes of arrows.

### Translations

In a plane, we can interpret the geometric vector  $q$ , comprising arrows lying in the plane, as a command to each point of the plane to move to a new position. For example, a particular point  $P$  goes to the point  $Q$  if  $\overrightarrow{PQ} \in q$ .



This is very reminiscent of the idea of a mapping in which  $Q$  is the image of  $P$ , and in fact this is sometimes a helpful way of looking at geometric vectors. We can use the geometric vector  $q$  to define a one-one function  $f$ , with domain and codomain the set of points in the plane, such that

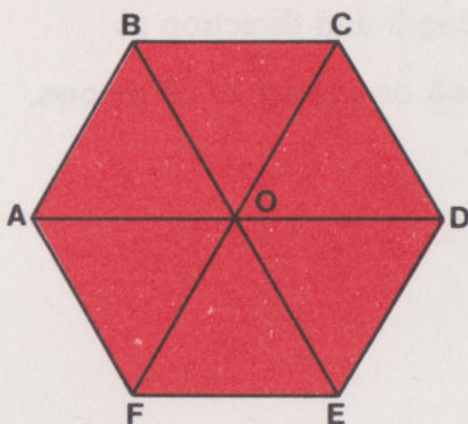
$$f: P \longmapsto Q$$

where  $\overrightarrow{PQ} = q$ . Such a function is called a **translation**. To each geometric vector there corresponds a unique translation.

In a similar way we could define a translation of three-dimensional space.



## Exercise 1



The above figure is a regular hexagon. We have omitted the arrow heads as these are implied in the following statements. In each case indicate if the statement is true or false:

- (i)  $\overrightarrow{AB} = \overrightarrow{ED}$
- (ii)  $\overrightarrow{AB} = \underline{AB}$
- (iii)  $\underline{FO} = \underline{ED}$
- (iv)  $\underline{AO} = \underline{EF}$
- (v)  $\underline{BC} = \underline{AD}$

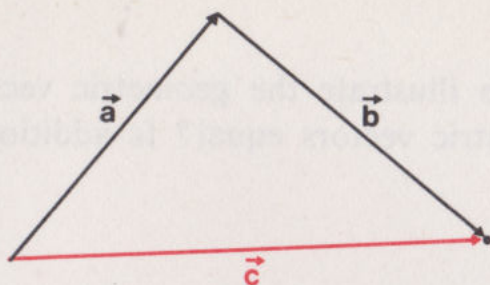
## 4.2 Addition on the Set of Geometric Vectors

We have our set of geometric vectors, and the problem is now to combine them in a mathematically interesting fashion. We might ask what sort of combinations will be useful, but since for the moment we are regarding this as a mathematical exercise rather than as an attack on an external problem, that question is not really relevant. If we find (as, of course, we do) that our mathematical system has some useful applications, so much the better, but that is not our prime concern at present.

First we shall define an operation  $+$  which we shall call *addition* of geometric vectors; *addition* is a good word to use because the operation has very similar properties to the operation of addition on  $R$ .

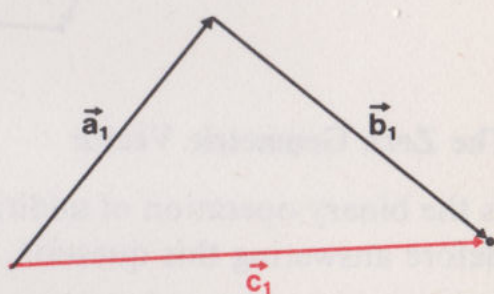
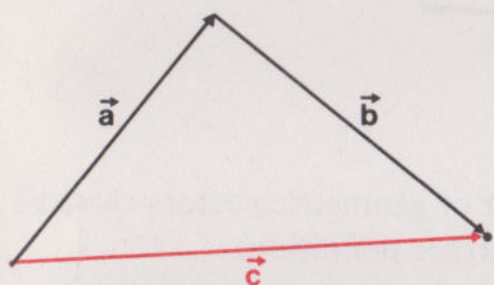
For any two geometric vectors  $\underline{a}$  and  $\underline{b}$  we define  $\underline{a} + \underline{b}$  as follows. Take any arrow  $\vec{a} \in \underline{a}$ , then choose the arrow  $\vec{b} \in \underline{b}$  which has its blunt end at the sharp end of  $\vec{a}$ .





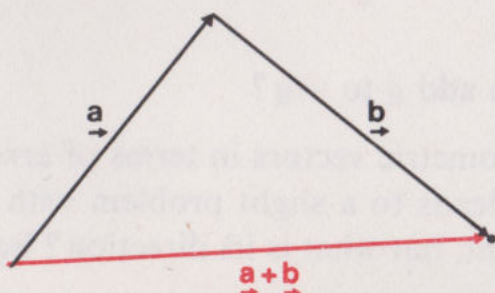
The arrow  $\vec{c}$ , with its blunt end at the blunt end of  $\vec{a}$  and sharp end at the sharp end of  $\vec{b}$ , belongs to some geometric vector  $\zeta$ , and we define this geometric vector to be  $\vec{a} + \vec{b}$ .

Have we really defined a binary operation on the set of geometric vectors? One of the requirements of a binary operation is that it should give a *unique* answer. We formed  $\vec{a} + \vec{b} = \zeta$  by taking any  $\vec{a} \in \mathfrak{a}$ , which determined  $\vec{b} \in \mathfrak{b}$ , and then combining  $\vec{a}$  and  $\vec{b}$  to give  $\zeta$ . Suppose we take a different representative  $\vec{a}_1 \in \mathfrak{a}$ ; will we still get the same  $\zeta$ ?



The two triangles shown are congruent, and it follows that  $\vec{c}$  and  $\vec{c}_1$  have the same length: also  $\vec{c}$  is parallel to  $\vec{c}_1$ , so  $\vec{c}$  and  $\vec{c}_1$  have the same direction. So both  $\vec{c}$  and  $\vec{c}_1$  belong to  $\zeta$ . The geometric vector  $\zeta$  is therefore uniquely defined.

We can represent  $\vec{a} + \vec{b}$  by the following diagram. (The arrows in this diagram are merely representatives from the corresponding geometric vectors; we have labelled them as geometric vectors.)





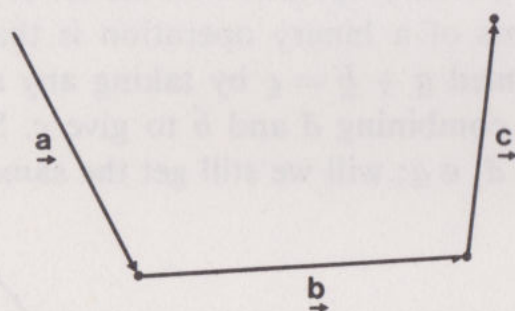
*Exercise 1*

Draw a diagram to illustrate the geometric vectors  $\underline{a} + \underline{b}$  and  $\underline{b} + \underline{a}$ . Are the two geometric vectors equal? Is addition of geometric vectors commutative?

*Exercise 2*

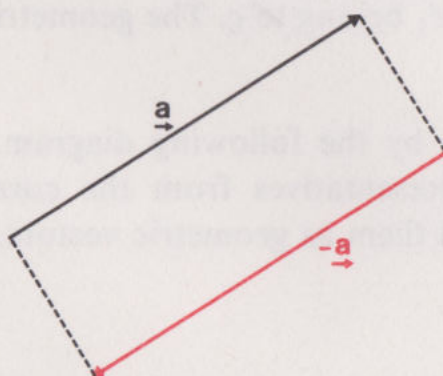
Use the following diagram to illustrate the associative property of addition of geometric vectors:

$$(\underline{a} + \underline{b}) + \underline{c} = \underline{a} + (\underline{b} + \underline{c})$$

**The Zero Geometric Vector**

Is the binary operation of addition on the set of geometric vectors closed? Before answering this question, we give a further definition.

Given a geometric vector  $\underline{a}$ , we define  $-\underline{a}$  to be the geometric vector determined by the arrow with the same length but the opposite direction to  $\underline{a}$ , where  $\underline{a} \in \mathcal{V}$ .



What happens if we add  $\underline{a}$  to  $-\underline{a}$ ?

We have defined geometric vectors in terms of arrows, which have length and direction; this leads to a slight problem with the “zero element”. A zero length is all right, but what is its direction? Because this question has



no satisfactory answer, we begin our process of abstraction and define the **zero geometric vector** as the result of adding *any* two geometric vectors of the form  $\vec{a}$  and  $-\vec{a}$ . Notice that  $\vec{a} + (-\vec{a})$  and  $\vec{b} + (-\vec{b})$  define the same zero geometric vector, which we denote by  $\vec{0}$ . (We are saying that, strictly speaking,  $\vec{0}$  is not a geometric vector, as it has no direction, but it is convenient to *call* it a geometric vector. In much the same way, the number zero in  $R$  is not really on a par with the other real numbers.) By analogy with zero in  $R$ , we define  $-\vec{0} = \vec{0}$  (since  $\vec{0}$  has zero length, we cannot “turn it round”).

It is consistent with our previous definitions to define

$$\vec{a} + \vec{0} = \vec{0} + \vec{a} = \vec{a}$$

(cf.  $a + 0 = 0 + a = a$  where  $a \in R$ ). The element  $\vec{0}$  corresponds to the translation which maps each point to itself, i.e. for which nothing moves.

With the inclusion of the zero geometric vector, our set of geometric vectors with addition is now closed.

### The Operation of Subtraction

Let us summarize our present position. We have the set of geometric vectors, with the closed binary operation of addition defined on it. We also have a zero element  $\vec{0}$  in the set. What about subtraction?

There is a difference in the real number system between

the element which when added to  $+2$  gives zero,

that is,

the signed integer  $-2$

and

the instruction: “subtract the number  $+2$ ”,

which we may write as

$$-(+2).$$

In the first case the  $-$  is a label attached to the 2 to show that the number is negative: in the second case the  $-$  denotes the binary operation of subtraction. For geometric vectors, we have the equivalent of the first case, but we do not yet have the equivalent of the second case as we have not defined *subtraction* of geometric vectors.

In the real number system, there is a connection between the signed

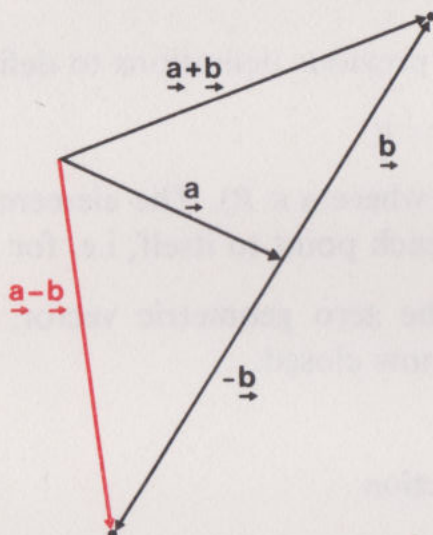


number  $-2$  and “subtract  $+2$ ”: “subtract  $+2$ ” is equivalent to “add  $-2$ ”.

By analogy, we now define

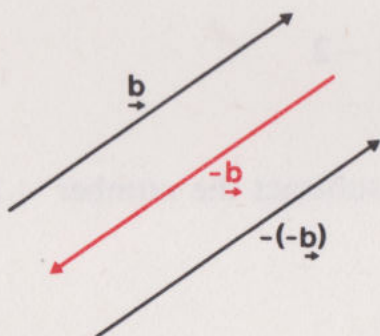
$$\underline{a} - \underline{b} \text{ to be } \underline{a} + (-\underline{b}).$$

This may seem obvious, but it is an essential step. The result of subtracting  $\underline{b}$  from  $\underline{a}$  is shown below.



Notice that the operation  $-$  is a binary operation. Also, from the definition of  $-\underline{b}$ , it follows that

$$-(-\underline{b}) = \underline{b}$$



So

$$\begin{aligned} \underline{a} - (-\underline{b}) &= \underline{a} + (-(-\underline{b})) \\ &= \underline{a} + \underline{b}. \end{aligned}$$



### Properties of Geometric Vectors

- (i)  $\underline{a} + \underline{b}$  is a geometric vector (+ is closed).
- (ii)  $\underline{a} + (\underline{b} + \underline{c}) = (\underline{a} + \underline{b}) + \underline{c}$  (+ is associative).
- (iii)  $\underline{a} + \underline{b} = \underline{b} + \underline{a}$  (+ is commutative).
- (iv) To each geometric vector  $\underline{a}$  there corresponds a unique geometric vector  $-\underline{a}$ .
- (v) There is a geometric vector  $\underline{0}$  such that for all  $\underline{a}$

$$\underline{a} + (-\underline{a}) = (-\underline{a}) + \underline{a} = \underline{0}$$

and

$$\underline{a} + \underline{0} = \underline{0} + \underline{a} = \underline{a}$$

- (vi)  $-\underline{0} = \underline{0}$ .
- (vii) Subtraction of geometric vectors is defined by

$$\underline{a} - \underline{b} = \underline{a} + (-\underline{b})$$

### Exercise 3

From the definitions, prove that

$$(\underline{a} - \underline{b} = \underline{c}) \text{ implies } (\underline{a} = \underline{c} + \underline{b})$$

## 4.3 Scalar Multiples of Geometric Vectors

In this section we consider lengths and scalar multiples of geometric vectors.

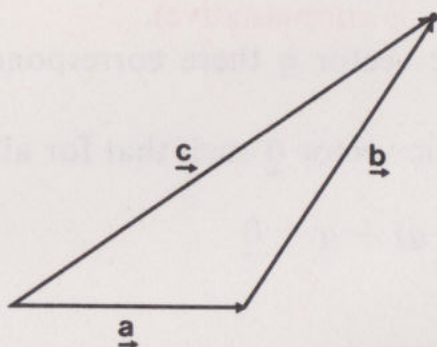
### Length of a Geometric Vector

It is convenient to have a notation for the length of a geometric vector, and we denote the **length of  $\underline{a}$**  by  $|\underline{a}|$  which we read as “the modulus of  $\underline{a}$ ”. We have used the word *modulus* previously for real numbers (e.g.  $|-2| = 2$ ), and we shall use it again for complex numbers. In each case where we use *modulus*, we are considering the magnitude of the quantity only, ignoring “direction”. (For example, on the real number line,  $-2$  and  $2$  are the same distance from  $0$ , but in opposite directions.) We write the **length of  $\underline{AB}$**  as either  $|\underline{AB}|$  or  $AB$ .



Suppose that

$$\underline{c} = \underline{a} + \underline{b}$$



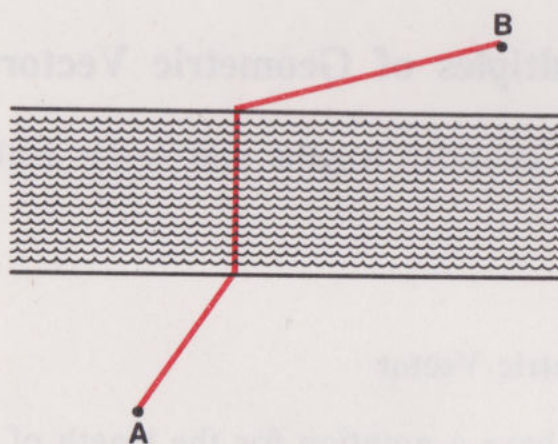
The length of any side of a triangle is less than (or equal to) the sum of the other two.\* In the modulus notation this statement is

$$|\underline{c}| \leq |\underline{a}| + |\underline{b}|$$

This inequality is called the **triangle inequality**.

It is only possible to have equality when the vertices of the triangle lie in a straight line, and then the interior of the triangle disappears completely.

### Example 1

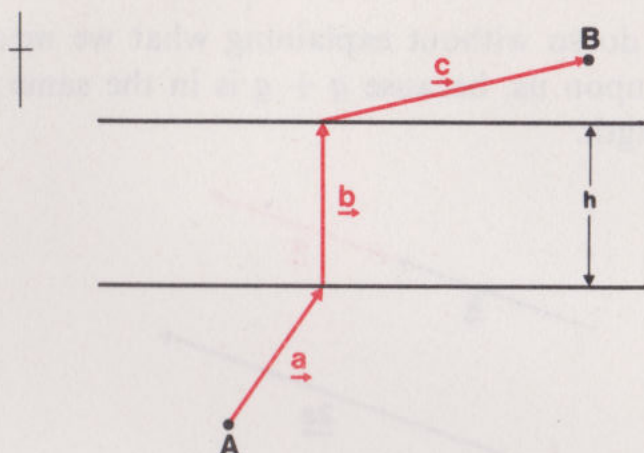


A (mythical) man wishes to travel regularly from  $A$  to  $B$ , and he wants to build a bridge across the river at right angles to the bank (to keep the bridge as short as possible) in the position which will minimize his travelling distance. Where should he put the bridge?

\* This is a theorem in Euclidean geometry.



It is a waste of time using calculus on this problem: there are easier ways of doing it. Consider the following diagram.



The geometric vectors represent the three distinct parts of the man's journey, and clearly he needs to walk in a straight line on each bank. Our problem is to choose the position of the bridge which will minimize

$$|\underline{a}| + |\underline{b}| + |\underline{c}|.$$

The value of  $|\underline{b}|$  is  $h$  and is not affected by the position of the bridge, so our problem is to minimize  $|\underline{a}| + |\underline{c}|$ . We know from the triangle inequality that

$$|\underline{a} + \underline{c}| \leq |\underline{a}| + |\underline{c}|,$$

and we only get equality when  $\underline{a}$  and  $\underline{c}$  are parallel.\* Hence, the minimum value occurs when  $\underline{a}$  and  $\underline{c}$  are parallel. It only remains to use this information to find the position of the bridge.

### Exercise 1

How does the man determine the position of the bridge to achieve the minimum distance?

### Multiplication of a Geometric Vector by a Scalar

We know what we mean by  $\underline{a} + \underline{a}$ , but if we were to abbreviate this to  $2\underline{a}$  without further ado, we would be missing a point.

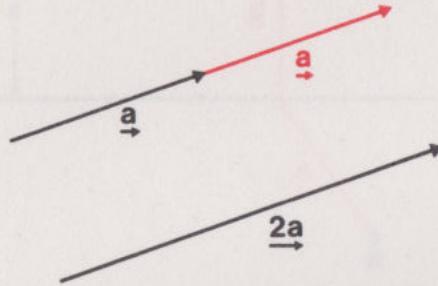
\* The geometric vectors  $\underline{a}$  and  $\underline{c}$  are defined to be parallel if the arrows  $\vec{a}$  and  $\vec{c}$  are parallel, where  $\vec{a} \in \underline{a}$  and  $\vec{c} \in \underline{c}$ .



It seems natural to write

$$\underline{a} + \underline{a} = 2\underline{a},$$

but we should not do so without explaining what we mean by  $2\underline{a}$ . The definition is thrust upon us, because  $\underline{a} + \underline{a}$  is in the same direction as  $\underline{a}$  but has twice its length.



For  $\lambda \in R$  we define  $\lambda \underline{a}$  ( $\lambda > 0$ ) to be a geometric vector in the same direction as  $\underline{a}$  but with length  $\lambda|\underline{a}|$ . This implies that  $|\lambda \underline{a}| = \lambda|\underline{a}|$ . If  $\lambda < 0$ , then we define  $\lambda \underline{a}$  to be in the opposite direction to  $\underline{a}$ , with length  $-\lambda|\underline{a}|$ . We define  $0\underline{a} = \underline{0}$ . So

$$|\lambda \underline{a}| = |\lambda| \times |\underline{a}|.$$

When we write  $\lambda \underline{a}$ , we say that the geometric vector  $\underline{a}$  is *multiplied* by the *scalar*  $\lambda$ . Although  $\lambda$  is just a real number in the present context, we call it a **scalar** to distinguish it from a geometric vector, because there exist more general situations in which  $\lambda$  is a scalar but not a real number.

### Further Properties of Geometric Vectors

Using our definition of  $\lambda \underline{a}$ , you should be able to prove the following results, which are in accord with our intuitive expectations:

- (viii) When  $\lambda = 0$ , we have  $0\underline{a} = \underline{0}$ .
- (ix) When  $\lambda = 1$ , we have  $1\underline{a} = \underline{a}$ .
- (x) When  $\lambda = -1$ , we have  $-1\underline{a} = -\underline{a}$ .
- (xi) For any geometric vector  $\underline{a}$  and any real number  $\lambda$ ,  $\lambda \underline{a}$  is a geometric vector.
- (xii)  $\lambda(\underline{a} + \underline{b}) = \lambda \underline{a} + \lambda \underline{b}$  for any geometric vectors  $\underline{a}$  and  $\underline{b}$  and any real number  $\lambda$  (see the following exercise).
- (xiii) For any geometric vector  $\underline{a}$  and any real numbers  $\lambda$  and  $\mu$ ,

$$(\lambda + \mu)\underline{a} = \lambda \underline{a} + \mu \underline{a}$$

- (xiv) For any geometric vector  $\underline{a}$  and any real numbers  $\lambda$  and  $\mu$ ,

$$(\lambda \mu)\underline{a} = \lambda(\mu \underline{a})$$



*Exercise 2*

Demonstrate property (xii), i.e. that the function  $q \mapsto \lambda q$  is a morphism of the set of geometric vectors under addition to itself:

$$\lambda(q + b) = \lambda q + \lambda b$$

The next exercise serves as an introduction to the following section in which we shall be interested in the following questions:

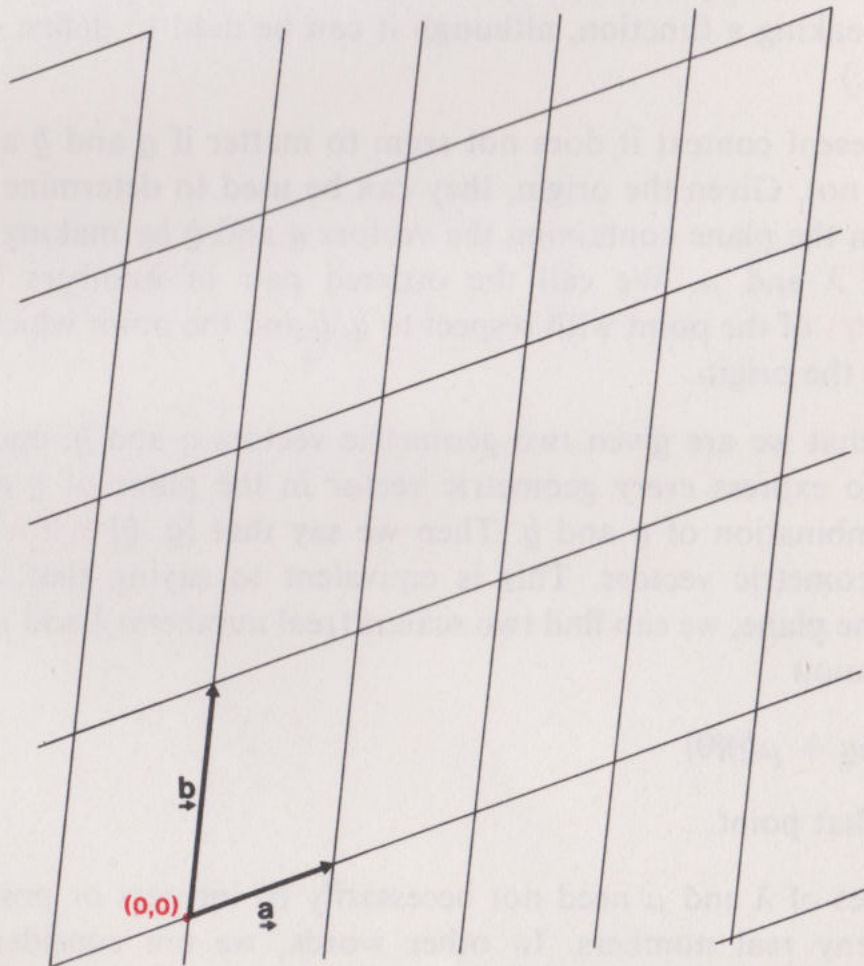
- (i) If we are given two geometric vectors  $q$  and  $b$ , can we add suitably chosen scalar multiples of them to produce a third given geometric vector  $c$ ?
- (ii) If we produce  $c$  as a sum of scalar multiples of  $q$  and  $b$ , is it possible to do it in several distinct ways?

*Exercise 3*

If  $q$  and  $b$  are the two geometric vectors represented in the following diagram, find the point determined as the image of the origin  $O$  under the translation corresponding to

$$\lambda q + \mu b,$$

where  $\lambda = 2$  and  $\mu = 3$ . Label this point  $(2, 3)$ .





In general, if  $f$  is the translation corresponding to  $\lambda \underline{a} + \mu \underline{b}$ , and  $f(0) = P$ , then we label  $P$   $(\lambda, \mu)$ .

Find the points on the above diagram which have labels  $(1, 1)$ ,  $(3, 2)$ ,  $(0, 1)$ ,  $(0, 0)$ ,  $(-1, 1)$ .

Now try the problem in reverse. For any point you like on the diagram, find (approximate) values of  $\lambda$  and  $\mu$  such that  $(\lambda, \mu)$  is the label for your chosen point.

## 4.4 Linear Dependence and Independence

The previous exercise is reminiscent of the familiar rectangular Cartesian co-ordinate system. In fact, had we chosen  $\underline{a}$  and  $\underline{b}$  at right angles and of unit length, then the situation would be identical, and the pair  $(\lambda, \mu)$  would be the co-ordinates of the point which is the image of the origin under the translation corresponding to  $\lambda \underline{a} + \mu \underline{b}$ . We call an expression such as  $\lambda \underline{a} + \mu \underline{b}$  a **linear combination of  $\underline{a}$  and  $\underline{b}$** .

(To save ourselves having to repeat the phrase “the image of the origin under the translation corresponding to  $\underline{a}$ ” we shall write  $\underline{a}(0)$ . This is an abuse of notation (a standard practice in mathematics) because  $\underline{a}$  is not strictly speaking a function, although it can be used to define one, as we have seen.)

In our present context it does not seem to matter if  $\underline{a}$  and  $\underline{b}$  are at right angles or not. Given the origin, they can be used to determine any point we wish in the plane containing the vectors  $\underline{a}$  and  $\underline{b}$  by making a suitable choice of  $\lambda$  and  $\mu$ . We call the ordered pair of numbers  $(\lambda, \mu)$  the **co-ordinates** of the point with respect to  $\underline{a}$ ,  $\underline{b}$  and the point which has been chosen as the origin.

Suppose that we are given two geometric vectors  $\underline{a}$  and  $\underline{b}$ , and that it is possible to express *every* geometric vector in the plane of  $\underline{a}$  and  $\underline{b}$  as a linear combination of  $\underline{a}$  and  $\underline{b}$ . Then we say that  $\{\underline{a}, \underline{b}\}$  **spans** the set of (plane) geometric vectors. This is equivalent to saying that, given any point of the plane, we can find two scalars (real numbers)  $\lambda$  and  $\mu$  such that the expression

$$(\lambda \underline{a} + \mu \underline{b})(0)$$

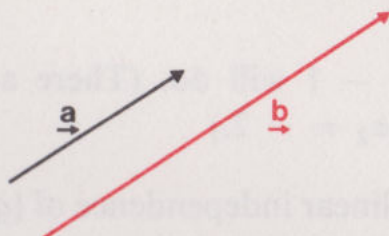
specifies that point.

(The values of  $\lambda$  and  $\mu$  need not necessarily be integers or positive; they may be any real numbers. In other words, we are considering every



point of the plane and not just those on the vertices of the grid as in Exercise 4.3.3.)

Will a set of any pair of geometric vectors span the set of (plane) geometric vectors? The answer is “No”. To see this, suppose that we are given the pair of geometric vectors shown in the following diagram.



Whatever values of  $\lambda$  and  $\mu$  we choose, the points  $(\lambda\vec{a} + \mu\vec{b})(0)$  always lie in a straight line: we cannot get out of that line with this choice of geometric vectors. Essentially this is because  $\vec{b}$  is a multiple of  $\vec{a}$ ; if  $\vec{b} = 2\vec{a}$ , for example, then

$$\begin{aligned}\lambda\vec{a} + \mu\vec{b} &= \lambda\vec{a} + \mu(2\vec{a}) \\ &= (\lambda + 2\mu)\vec{a},\end{aligned}$$

and every linear combination of  $\vec{a}$  and  $\vec{b}$  is simply a multiple of  $\vec{a}$ .

In this case  $\{\vec{a}, \vec{b}\}$  only spans the set of geometric vectors parallel to  $\vec{a}$ , and this is because  $\vec{b}$  contributes nothing new.

Things will go wrong (i.e.  $\{\vec{a}, \vec{b}\}$  will not span the set of plane geometric vectors) if  $\vec{b}$  is a multiple of  $\vec{a}$ , that is to say, if  $\vec{b} = \beta\vec{a}$  for some real number  $\beta$  (including  $\beta = 0$ ). Equally well, we are in trouble if  $\vec{a}$  is a multiple of  $\vec{b}$ , so that  $\vec{a} = \alpha\vec{b}$  for some real number  $\alpha$ . This can be expressed in a symmetric form: things will go wrong if we can find two real numbers  $\alpha_1$  and  $\alpha_2$ , which are *not both zero*, such that

$$\alpha_1\vec{a} + \alpha_2\vec{b} = \vec{0}$$

We say that  $\{\vec{a}, \vec{b}\}$  is *linearly dependent* if and only if there are scalars  $\alpha_1$  and  $\alpha_2$ , not both zero, which satisfy the above equation. We say that  $\{\vec{a}, \vec{b}\}$  is *linearly independent* if it is not linearly dependent.

There is a simple, but important, consequence of linear independence. If we are told that  $\{\vec{a}, \vec{b}\}$  is linearly independent, and yet there *are* scalars  $\alpha_1$  and  $\alpha_2$  such that

$$\alpha_1\vec{a} + \alpha_2\vec{b} = \vec{0}$$

then we can immediately deduce that  $\alpha_1 = \alpha_2 = 0$ .



*Example 1*

We can show that  $\{\underline{a}, 2\underline{a}\}$  is linearly dependent for any geometric vector  $\underline{a}$  as follows.

We need to find two scalars  $\alpha_1$  and  $\alpha_2$ , not both zero, such that

$$\alpha_1 \underline{a} + \alpha_2 (2\underline{a}) = \underline{0}$$

Clearly  $\alpha_1 = 2$  and  $\alpha_2 = -1$  will do. (There are many other choices; for example,  $\alpha_1 = 4$  and  $\alpha_2 = -2$ .)

We shall see later that the linear independence of  $\{\underline{a}, \underline{b}\}$  is not only *necessary* but also *sufficient* to guarantee that it spans the set of (plane) geometric vectors.

Very similar arguments apply in three-dimensional space. We can extend the same idea to any number of geometric vectors to give the following definitions.

The set of geometric vectors  $\{\underline{q}^{(1)}, \underline{q}^{(2)}, \underline{q}^{(3)}, \dots, \underline{q}^{(n)}\}$  is said to be **linearly dependent** if and only if there are scalars  $\alpha_1, \alpha_2, \dots, \alpha_n$ , not all zero, such that

$$\alpha_1 \underline{q}^{(1)} + \alpha_2 \underline{q}^{(2)} + \alpha_3 \underline{q}^{(3)} + \dots + \alpha_n \underline{q}^{(n)} = \underline{0}.$$

(We shall also say that the geometric vectors themselves are linearly dependent.)

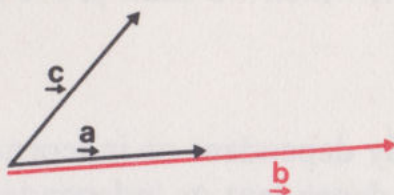
The set of geometric vectors  $\{\underline{q}^{(1)}, \underline{q}^{(2)}, \dots, \underline{q}^{(n)}\}$  is said to be **linearly independent** if it is not linearly dependent. (We shall also say that the geometric vectors themselves are linearly independent.)

The essential feature about a linearly dependent set of geometric vectors is that it implies a certain amount of redundancy in the set under consideration. If the set is linearly independent, there is no such redundancy although it may still be “incomplete”; for example, a linearly independent set  $\{\underline{a}, \underline{b}\}$  cannot span the whole set of geometric vectors in three dimensions. To make this intuitive idea clearer, let us look again at the two-dimensional problem, not this time in terms of two geometric vectors  $\underline{a}$  and  $\underline{b}$  lying in the plane, but in terms of three geometric vectors  $\underline{a}$ ,  $\underline{b}$  and  $\underline{c}$  lying in the plane. There are two possible occurrences. Two or more of the geometric vectors are parallel, or they all have different directions (assuming that none of them is  $\underline{0}$ ). Let us consider the two cases.

(i) Suppose first that  $\underline{a}$  and  $\underline{b}$  are parallel, then  $\{\underline{a}, \underline{b}\}$  is linearly dependent,

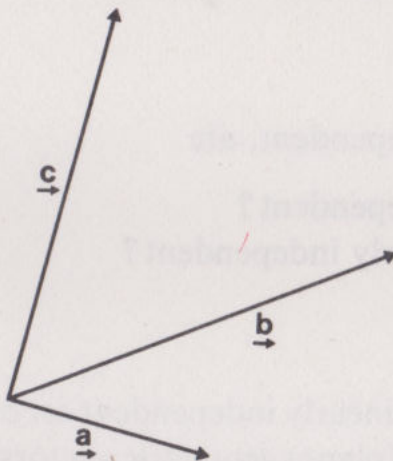


because we can express one of the geometric vectors as a scalar multiple of the other:  $\underline{a} = \lambda \underline{b}$  say.



This means that  $\underline{a}$  is effectively redundant, and everywhere it occurs we can simply replace it by  $\lambda \underline{b}$ .

- (ii) Now suppose that  $\underline{a}$  and  $\underline{b}$  are not parallel, then  $\{\underline{a}, \underline{b}\}$  is linearly independent, so that *intuitively* we feel sure that it spans the plane.



In this case  $\underline{c}$  can be expressed in the form  $\lambda \underline{a} + \mu \underline{b}$ , so that

$$\underline{c} = \lambda \underline{a} + \mu \underline{b},$$

and everywhere  $\underline{c}$  occurs it can be replaced by this linear combination of  $\underline{a}$  and  $\underline{b}$ . Here we have made  $\underline{c}$  the redundant element.

Notice that in the first case we have  $\underline{a} = \lambda \underline{b}$ , so that

$$\underline{a} - \lambda \underline{b} + 0\underline{c} = \underline{0},$$

and in the second case  $\underline{c} = \lambda \underline{a} + \mu \underline{b}$ , so that

$$\underline{c} - \lambda \underline{a} - \mu \underline{b} = \underline{0}$$

From the definition of linear dependence we can see that in both cases the set of geometric vectors  $\{\underline{a}, \underline{b}, \underline{c}\}$  is linearly dependent.

It appears that *any* set of three geometric vectors lying in a plane *must* be linearly dependent, and that a linearly independent set of just two



geometric vectors is needed to span the set of (plane) geometric vectors. The above argument does not constitute a proof, but our intuitive discussion suggests the result which we shall prove later.

### Exercise 1

- (i) Are  $\underline{a}$  and  $-\underline{a}$  linearly dependent or independent?
- (ii) Are  $\underline{a}$  and  $\underline{0}$  linearly dependent or independent?

### Exercise 2

If  $\underline{a}$  and  $\underline{b}$  are linearly dependent, are

- (i)  $3\underline{a}$  and  $4\underline{b}$  linearly dependent?
- (ii)  $\underline{a} + \underline{b}$  and  $\underline{a} - \underline{b}$  linearly dependent?

### Exercise 3

If  $\underline{a}$  and  $\underline{b}$  are linearly independent, are

- (i)  $3\underline{a}$  and  $4\underline{b}$  linearly independent?
- (ii)  $\underline{a} + \underline{b}$  and  $\underline{a} - \underline{b}$  linearly independent?

## Base Geometric Vectors

We shall see later that *any* linearly independent set of two geometric vectors does in fact span the set of (plane) geometric vectors; this leads us naturally to the following definition.

If a subset of a set of geometric vectors spans the whole set, and if in addition the subset is linearly independent, then we say that the subset forms a **basis** for the set. The elements of the subset are called **base geometric vectors**, or more briefly **base vectors**.

It seems clear that there will be two geometric vectors in a basis for the set of plane geometric vectors, and very likely that there will be three geometric vectors in a basis for the set of geometric vectors in three dimensions. We haven't *proved* either of these statements, and there are several vital things to be cleared up. For example, how do we know that *every* basis for a particular set of geometric vectors contains the same number of elements? We shall return to these questions later.

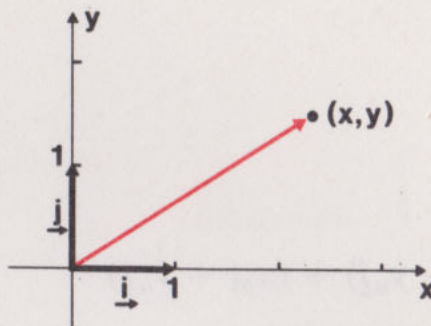
## 4.5 An Algebra of Number Pairs

In this section we shall look at the relationship between Cartesian co-ordinates and geometric vectors.



### Cartesian Co-ordinates and Geometric Vectors

We choose geometric vectors  $\underline{i}$  and  $\underline{j}$  of unit length, in the directions of the Cartesian  $x$  and  $y$  axes respectively.



For any point  $q(0)$  with co-ordinates  $(x, y)$  we have

$$q(0) = (x\underline{i} + y\underline{j})(0)$$

so that

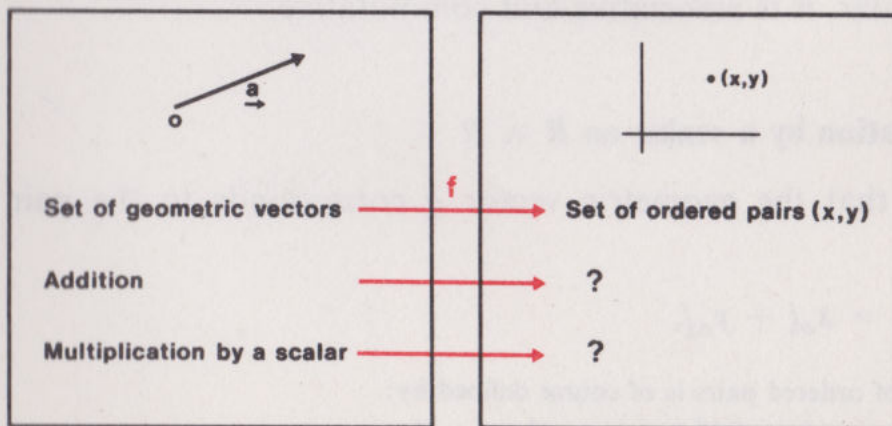
$$q = x\underline{i} + y\underline{j}$$

We have concentrated on geometric vectors, but, since there is a one-one relationship between geometric vectors in a plane and number pairs, we can also construct an *algebra of number pairs*.

Let us start then with the set of all ordered pairs of real numbers,  $R \times R$ , as the set on which the algebra is to be constructed. We shall call  $f$  the mapping from (plane) geometric vectors to rectangular Cartesian co-ordinates  $(x, y)$ . That is,

$$f: x\underline{i} + y\underline{j} \longmapsto (x, y)$$

We know that  $f$  is one-one, so operations on the set of geometric vectors will “carry over” to the set  $R \times R$ . What happens to addition and multiplication by a scalar under  $f$ ?





**Addition on  $R \times R$** 

Suppose that we are given two geometric vectors  $\underline{a}$  and  $\underline{b}$  corresponding to the pairs  $(x_a, y_a)$  and  $(x_b, y_b)$  respectively. Then

$$\underline{a} = x_a \underline{i} + y_a \underline{j}$$

and

$$\underline{b} = x_b \underline{i} + y_b \underline{j}$$

It follows that

$$\begin{aligned} \underline{a} + \underline{b} &= (x_a \underline{i} + y_a \underline{j}) + (x_b \underline{i} + y_b \underline{j}) \\ &= (x_a + x_b) \underline{i} + (y_a + y_b) \underline{j}, \end{aligned}$$

using the associativity and commutativity of addition on the set of geometric vectors. The sum  $\underline{a} + \underline{b}$  therefore corresponds to the pair  $(x_a + x_b, y_a + y_b)$  and we are led to define the induced addition on  $R \times R$  by\*

$$(x_a, y_a) + (x_b, y_b) = (x_a + x_b, y_a + y_b)$$

We have, as it were, used the black arrows in the following commutative diagram to define the  $+$ .

$$\begin{array}{ccc} (a, b) & \xrightarrow{+} & a + b \\ \downarrow f & & \downarrow f \\ ((x_a, y_a), (x_b, y_b)) & \xrightarrow{+} & (x_a + x_b, y_a + y_b) \end{array}$$

Notice that the two  $+$ 's are different, being defined on different sets, but because of the way  $+$  has been defined as the induced binary operation on  $R \times R$ , it has the same properties as  $+$  on the set of geometric vectors. In particular, it is associative and commutative.

**Multiplication by a scalar on  $R \times R$** 

Suppose that the geometric vector  $\underline{a}$  corresponds to the pair  $(x_a, y_a)$ ; then

$$\underline{a} = x_a \underline{i} + y_a \underline{j},$$

\* Equality of ordered pairs is of course defined by:

$(x, y) = (x', y')$  if  $x = x'$  and  $y = y'$ .



so that

$$\begin{aligned}\lambda \underline{q} &= \lambda(x_a \underline{i} + y_a \underline{j}) \\ &= \lambda x_a \underline{i} + \lambda y_a \underline{j} \quad (\text{by property (xii) on p. 122})\end{aligned}$$

The geometric vector  $\lambda \underline{q}$  therefore corresponds to the ordered pair  $(\lambda x_a, \lambda y_a)$ , and we are led to define **multiplication by a scalar** on  $R \times R$  by

$$\lambda(x_a, y_a) = (\lambda x_a, \lambda y_a)$$

We have the following commutative diagram.

$$\begin{array}{ccc} \underline{q} & \xrightarrow{\times \lambda} & \lambda \underline{q} \\ \downarrow f & & \downarrow f \\ (x_a, y_a) & \xrightarrow{\times \lambda} & (\lambda x_a, \lambda y_a) \end{array}$$

We can, of course, define subtraction by analogy, also retaining all the properties that subtraction has for the set of geometric vectors. We are now able to leave the algebra of geometric vectors if we wish, and concentrate on the algebra of number pairs. But why restrict ourselves to number pairs? Why not triples, or  $n$ -tuples for that matter? In fact, we need not even restrict ourselves to any of these. Why not discuss some *abstract system* which has the essential properties of the above system? For then any results which we establish *apply to any such system*. We shall consider this later.

## 4.6 “Multiplication” on the Set of Geometric Vectors

We shall work in three-dimensional space, because it is here that the concepts are really useful.

In our construction of an algebra on the set of geometric vectors we have so far avoided the problem of defining a binary operation which could correspond to the operation of multiplication on the set of real numbers. It is true that we have defined multiplication by a scalar, but this does not combine two elements from the set of geometric vectors, but simply a geometric vector with a scalar.

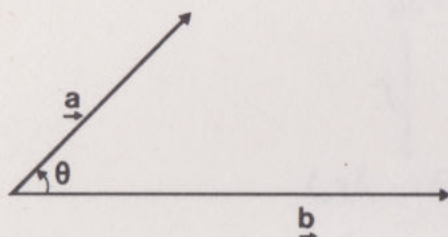
There are in fact several different operations which we can define, all having some features in common with multiplication of real numbers. The criteria for any particular choice of operation can be varied, but one criterion high on the list must be usefulness, either within mathematics or



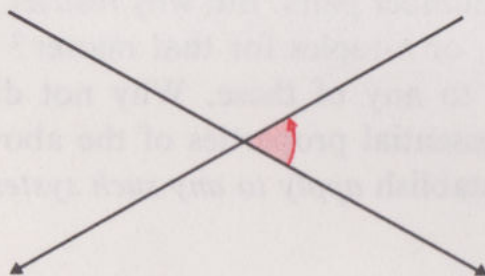
in the application of mathematics. We shall choose a particular operation, and we shall demonstrate one of its uses later. (See also Exercise 1.) The chosen operation is called the *inner product*, sometimes called the *scalar product* or *dot product*.

### The Inner Product

Suppose that we are given two geometric vectors  $\underline{a}$  and  $\underline{b}$  with an angle  $\theta$  between them (i.e.  $\theta$  is the angle between any two arrows  $\vec{a}$  and  $\vec{b}$ , where  $\vec{a} \in \underline{a}$  and  $\vec{b} \in \underline{b}$ ).

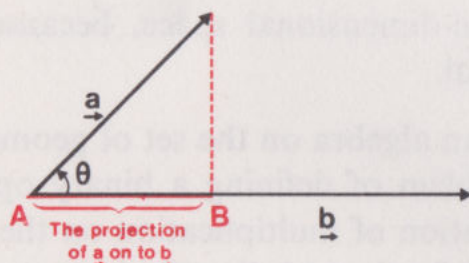


Notice that the angle  $\theta$  lies between the two arrow heads, *not* like the angle in the following diagram; also, we take  $\theta$  such that  $0 \leq \theta \leq \pi$ .



The *inner product* of  $\underline{a}$  and  $\underline{b}$ , denoted by  $\underline{a} \cdot \underline{b}$  is defined by

$$\underline{a} \cdot \underline{b} = |\underline{a}| |\underline{b}| \cos \theta$$



The length  $AB$  is called the *projection of  $\underline{a}$  on to  $\underline{b}$*  and is equal to  $|\underline{a}| \cos \theta$ . So the inner product of  $\underline{a}$  and  $\underline{b}$  is equal to  $|\underline{b}|$  times the projection of  $\underline{a}$  on to  $\underline{b}$ , and it is also equal to  $|\underline{a}|$  times the projection of  $\underline{b}$  on to  $\underline{a}$ . Notice that this operation is a *commutative binary operation* (just like multiplication of real numbers) but it is *not closed* on the set of geometric vectors, because the combination of two geometric vectors is *not* a geometric



vector but a scalar (a real number). This means, among other things, that we cannot define an “inverse” operation adequately, i.e. that there is no possibility of finding an equivalent of division. If we consider what we mean by division on the set of real numbers, we see that it is an operation which “undoes” the work of multiplication. Thus, if we take a real number  $a$  and multiply it by a non-zero number  $b$ , then divide the result by  $b$ , we are back to  $a$ . But if we take a geometric vector  $\underline{a}$  and form its inner product with a geometric vector  $\underline{b}$ , then we get  $\underline{a} \cdot \underline{b}$  which is a real number, and the problem of getting back from this real number to the geometric vector  $\underline{a}$  is not analogous to the multiplication/division problem.

This operation is very useful however. You should notice particularly that:

- (i) if  $\theta$  is the angle between two non-zero geometric vectors  $\underline{a}$  and  $\underline{b}$ , then

$$\cos \theta = \frac{\underline{a} \cdot \underline{b}}{|\underline{a}| |\underline{b}|};$$

- (ii)  $(\underline{a} \cdot \underline{b} = 0)$  if and only if ( $\underline{a}$  and  $\underline{b}$  are perpendicular)

i.e.  $\underline{a} \cdot \underline{b} = 0$  if and only if  $\cos \theta = 0$

provided, of course, that neither  $\underline{a}$  nor  $\underline{b}$  is the zero geometric vector;

- (iii)  $\underline{a} \cdot \underline{a} = |\underline{a}|^2$ , since  $\cos 0 = 1$ .

So if we can find a convenient way of calculating the inner product (as we can) which does not directly involve the definition given above, then we can use the inner product:

- (i) to calculate the angle between two geometric vectors;
- (ii) to determine whether or not two geometric vectors are perpendicular;
- (iii) to calculate the length of a geometric vector.

What properties of ordinary multiplication of real numbers does the inner product have?

It follows from the definition of inner product that

$$\underline{a} \cdot \underline{b} = \underline{b} \cdot \underline{a}, \text{ i.e. the inner product is commutative}$$

$$\lambda \underline{a} \cdot \mu \underline{b} = \lambda \mu \underline{a} \cdot \underline{b}.$$

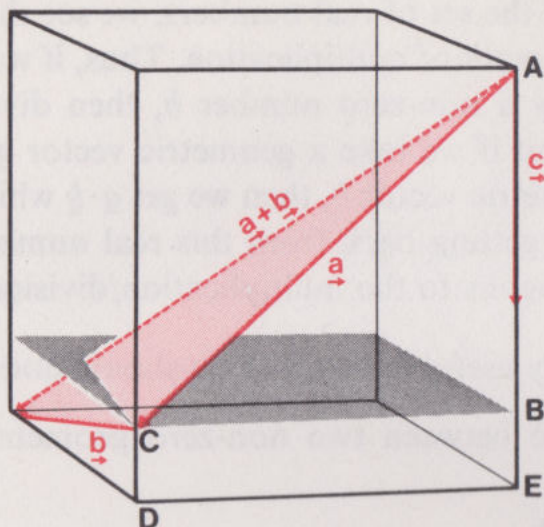
Associativity does not apply, since the inner product is not closed.\*

Is the inner product distributive over addition of geometric vectors?

\* We cannot form  $(\underline{a} \cdot \underline{b}) \cdot \underline{c}$  because  $(\underline{a} \cdot \underline{b})$  is a *scalar*, not a geometric vector, and we therefore cannot form its inner product with  $\underline{c}$ , i.e.  $(\underline{a} \cdot \underline{b}) \cdot \underline{c}$  is *meaningless*.



Consider the three arbitrary geometric vectors (in three dimensions)  $\underline{a}$ ,  $\underline{b}$  and  $\underline{c}$  illustrated in the following diagram:



From the diagram we can see that

$$\underline{AE} = \underline{AB} + \underline{CD},$$

so that

$$\begin{aligned} &(\text{the projection of } (\underline{a} + \underline{b}) \text{ on to } \underline{c}) = \\ &(\text{the projection of } \underline{a} \text{ on to } \underline{c}) + (\text{the projection of } \underline{b} \text{ on to } \underline{c}). \end{aligned}$$

Multiplying both sides of the equation by  $|\underline{c}|$ , we get

$$(\underline{a} + \underline{b}) \cdot \underline{c} = \underline{a} \cdot \underline{c} + \underline{b} \cdot \underline{c}$$

i.e. the inner product is right-distributive over addition of geometric vectors.

### Exercise 1

Use the distributive result above and the commutative property of the inner product to show that

$$\underline{c} \cdot (\underline{a} + \underline{b}) = \underline{c} \cdot \underline{a} + \underline{c} \cdot \underline{b},$$

i.e. that  $\cdot$  is left-distributive over  $+$ .

The result of Exercise 1 proves that the inner product is distributive over addition of geometric vectors.

We have seen that the inner product has some properties in common with ordinary multiplication, but there are important differences, essentially because the inner product is not closed. Although we have used a dot for



the inner product, and we often use a dot for multiplication of real numbers, we must be careful not to make unwarranted deductions. For example, the statement

$$(\underline{a} \cdot \underline{b} = 0) \text{ implies } (\underline{a} = \underline{0}) \text{ or } (\underline{b} = \underline{0})$$

is FALSE, because  $\underline{a} \cdot \underline{b} = 0$  also if neither  $\underline{a}$  nor  $\underline{b}$  is  $\underline{0}$  but  $\underline{a}$  is perpendicular to  $\underline{b}$ . Again the statement

$$(\underline{a} \cdot \underline{b} = \underline{a} \cdot \underline{c}) \text{ implies } (\underline{b} = \underline{c})$$

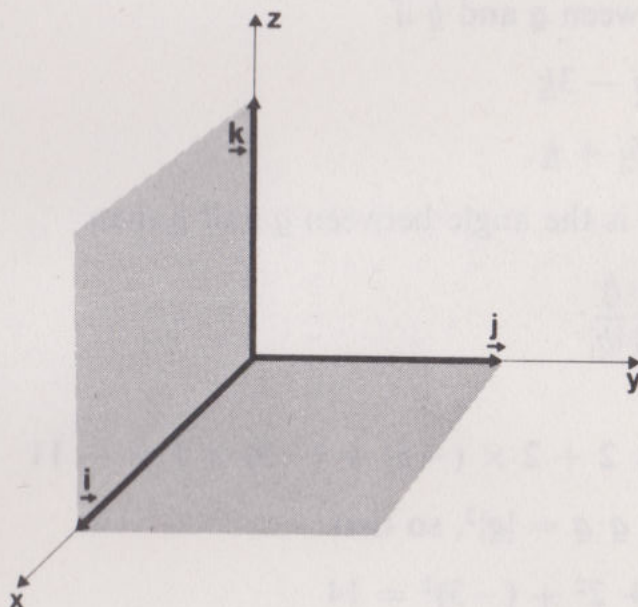
is also FALSE, because

$$|\underline{a}| |\underline{b}| \cos \theta_1 = |\underline{a}| |\underline{c}| \cos \theta_2,$$

where  $\theta_1, \theta_2$  are the angles between  $\underline{a}$  and  $\underline{b}$ ,  $\underline{a}$  and  $\underline{c}$  respectively, does not imply that

$$\underline{b} = \underline{c}.$$

The inner product is particularly easy to deal with because of the simple form it takes when the geometric vectors are written in terms of the basis  $\{\underline{i}, \underline{j}, \underline{k}\}$  consisting of geometric vectors of unit length in the directions of the  $x$ ,  $y$  and  $z$  Cartesian axes.



We have:

$$|\underline{i}| = |\underline{j}| = |\underline{k}| = 1;$$

the angle between any base vector and itself is zero;

the angle between any two distinct base vectors is  $\frac{\pi}{2}$ .



From the definition of the inner product we can see that

$$\begin{aligned} \underline{i} \cdot \underline{i} &= 1 & \text{and} & & \underline{i} \cdot \underline{j} &= \underline{j} \cdot \underline{i} &= 0 \\ \underline{j} \cdot \underline{j} &= 1 & & & \underline{j} \cdot \underline{k} &= \underline{k} \cdot \underline{j} &= 0 \\ \underline{k} \cdot \underline{k} &= 1 & & & \underline{k} \cdot \underline{i} &= \underline{i} \cdot \underline{k} &= 0 \end{aligned}$$

Suppose now that

$$\underline{a} = x_a \underline{i} + y_a \underline{j} + z_a \underline{k}$$

and

$$\underline{b} = x_b \underline{i} + y_b \underline{j} + z_b \underline{k}.$$

Using the above results for the inner products involving  $\underline{i}$ ,  $\underline{j}$  and  $\underline{k}$ , and the distributive law, we obtain

$$\begin{aligned} (\underline{x}_a \underline{i} + \underline{y}_a \underline{j} + \underline{z}_a \underline{k}) \cdot (\underline{x}_b \underline{i} + \underline{y}_b \underline{j} + \underline{z}_b \underline{k}) \\ = x_a x_b + y_a y_b + z_a z_b, \end{aligned}$$

so that we obtain the important formula

$$\underline{a} \cdot \underline{b} = x_a x_b + y_a y_b + z_a z_b$$

### Example 1

Find the angle between  $\underline{a}$  and  $\underline{b}$  if

$$\underline{a} = \underline{i} + 2\underline{j} - 3\underline{k}$$

and  $\underline{b} = 2\underline{i} - 5\underline{j} + \underline{k}.$

We know that if  $\theta$  is the angle between  $\underline{a}$  and  $\underline{b}$  then

$$\cos \theta = \frac{\underline{a} \cdot \underline{b}}{|\underline{a}| |\underline{b}|}.$$

Now

$$\underline{a} \cdot \underline{b} = 1 \times 2 + 2 \times (-5) + (-3) \times 1 = -11$$

and we know that  $\underline{a} \cdot \underline{a} = |\underline{a}|^2$ , so that

$$\begin{aligned} |\underline{a}|^2 &= 1^2 + 2^2 + (-3)^2 = 14 \\ |\underline{b}|^2 &= 2^2 + (-5)^2 + 1^2 = 30. \end{aligned}$$

Hence

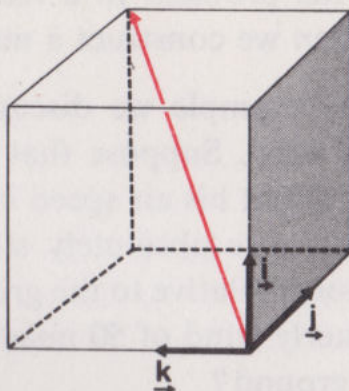
$$\cos \theta = \frac{-11}{\sqrt{14}\sqrt{30}}$$

Since  $\cos \theta$  is negative,  $\frac{\pi}{2} \leq \theta \leq \pi$ , and in fact  $\theta$  is approximately  $132^\circ$ .



*Example 2*

Find the angle between a diagonal of a cube and one of its edges.



We choose unit vectors  $\underline{i}$ ,  $\underline{j}$ ,  $\underline{k}$  in the directions of the edges of the cube, then the geometric vector  $\underline{q} = \underline{i} + \underline{j} + \underline{k}$  has the same direction as a diagonal of the cube, and

$$\underline{q} \cdot \underline{i} = |\underline{q}| \cos \theta$$

where  $\theta$  is the acute angle we wish to determine.

$$\underline{q} \cdot \underline{i} = (\underline{i} + \underline{j} + \underline{k}) \cdot \underline{i} = 1$$

also

$$|\underline{q}|^2 = \underline{q} \cdot \underline{q} = 3$$

so that  $\sqrt{3} \cos \theta = 1$ , i.e.  $\cos \theta = \frac{1}{\sqrt{3}}$ , so that  $\theta$  is approximately  $55^\circ$ .

## 4.7 Applications of Geometric Vectors

Our intention in this set of three volumes is to discuss certain topics in elementary mathematics which could be useful to the non-mathematician and certain other topics which are generally useful for learning and understanding any particular bit of mathematics. We never intended to discuss mathematics in application in any comprehensive way, but we have, from time to time discussed a particular application as an illustration. The final section of this chapter should be seen in this light; it is merely an indication of the beginnings of a very large subject: the application of geometric vectors.

The applied mathematician uses geometric vectors in essentially two distinct ways. First he makes use of them when constructing a mathematical model of a physical quantity such as *force*; and, secondly, he uses them as a convenient means for specifying the position of a point in space.

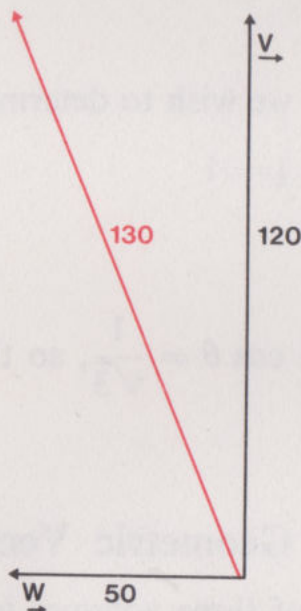


### Mathematical Modelling

The first stage of a problem in applied mathematics is usually one of simplification. We reduce the problem to a reasonable form by making certain assumptions, and then we construct a mathematical model.

This brings us back to the example we discussed in section 4.1 of an aeroplane flying in a cross wind. Suppose that the pilot points the nose of the aircraft due north and that his air speed indicator reads 120 mile/h. (This means that, if the air were absolutely still, the aircraft would be moving at 120 mile/h due north relative to the ground.) Let us also suppose that there is a constant easterly wind of 50 mile/h. What is the velocity of the aircraft relative to the ground?

Let us model the velocity of the aircraft relative to the wind by a geometric vector  $\underline{V}$ ; then we could represent the physical situation by the following diagram.



We have our model, but can we draw any conclusions? As the pilot tries to fly his aircraft due north it is constantly carried to the west at a rate of 50 mile/h. The resulting velocity relative to the ground can be modelled by the sum  $\underline{V} + \underline{W}$ , which in this case implies that the aircraft has a speed of 130 mile/h relative to the ground in the direction indicated on the diagram.

There is one extremely important point to notice about this example. We can model the individual velocities by geometric vectors, but in order to draw our final conclusion we need to assume that the addition operation for geometric vectors is the appropriate model of physical combinations of velocities.

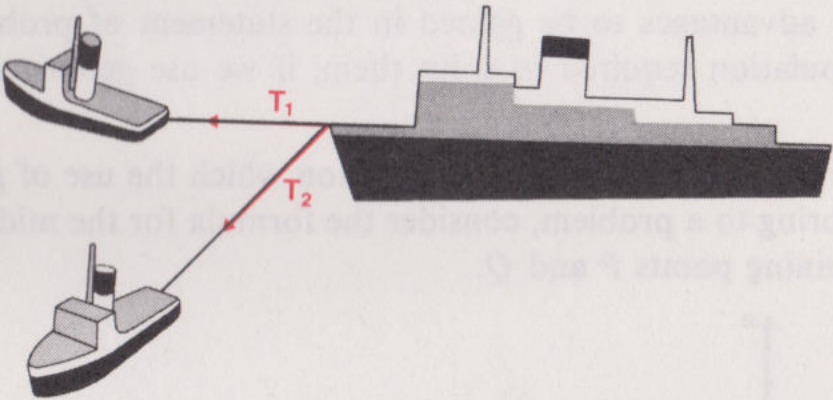


Physical Situation	Mathematical Model
Velocity of the wind	$\underline{W}$
Velocity of aircraft	$\underline{V}$
Combination of velocities	$+$

This, of course, depends on experimental verification or valid deduction from previously validated models. In this case the model apparently works, and so the physical and mathematical situations are related by a morphism.

Forces

Geometric vectors are often used to model *forces*. Suppose, for example, that two tug-boats are towing a ship.



We might make the gross simplification that the ship is a particle, and then represent the two forces in the tow-ropes by geometric vectors. What is the resulting force? The important point is that we can verify by experiment that forces on a particle are combined in the same way as geometric vectors are combined by addition. In other words, geometric vectors are an adequate representation in that their rule of combination corresponds to the physical combination of the quantities which they represent. To find the resulting force we simply *add* the geometric vectors, and this gives us the appropriate model of the net force.

It isn't always appropriate to model forces by geometric vectors. For example, if we wish to take account of the turning effect of the tow-ropes



on the ship, then the points at which they are attached will clearly be important. In this case we might simplify the ship not to a particle, but to a line segment, and we might model the forces by arrows attached to the appropriate points on the line segment.

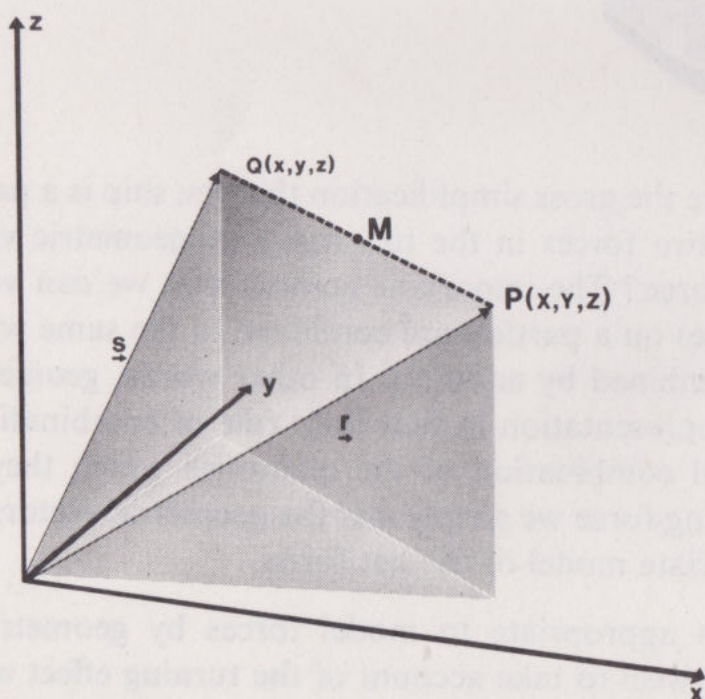
### Geometric Applications

The second important application of geometric vectors is their ability to specify points in space relative to a fixed point (called the origin). If we use a geometric vector  $\underline{r}$  to define a translation, then the image of the origin  $\underline{r}(O)$  is uniquely determined. This is the point  $P$  say.

It may be rather difficult at this stage to see the advantages of determining the point  $P$  by the geometric vector  $\underline{r}$  rather than by, say, its co-ordinates  $(x, y, z)$  in a Cartesian co-ordinate system. (See figure below.)

Often we wish to determine the position of a point in space in a problem which has arisen from a physical situation, and in which we have used geometric vectors to model quantities such as force. There are definite advantages in keeping all our discussion in terms of geometric vectors in this case. We could manage without geometric vectors, just as we could manage without (school) algebra and use arithmetic only. But there are considerable advantages to be gained in the statement of problems and in the manipulation required to solve them, if we use geometric vectors throughout.

As a very trivial example of the simplification which the use of geometric vectors can bring to a problem, consider the formula for the mid-point  $M$  of the line joining points  $P$  and  $Q$ .



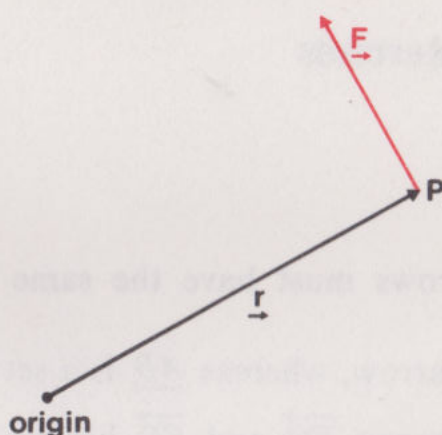


The Cartesian co-ordinates of  $M$  are

$$\left( \frac{x+X}{2}, \frac{y+Y}{2}, \frac{z+Z}{2} \right)$$

whereas the point is equally well determined by the geometric vector  $\frac{1}{2}(\underline{r} + \underline{s})$ . We require a third of the time and space to convey exactly the *same* information if we use geometric vectors.

Think for a few moments how you would convey the information on the following diagram (in which  $\underline{r}$  determines the point  $P$ , and  $\underline{F}$  is used to model a force applied to a particle at  $P$ ) in terms of Cartesian co-ordinates only.



## 4.8 Additional Exercises

### Exercise 1

We saw on page 113 that a geometric vector determines a unique translation which maps the set of points in the plane (or three-dimensional space) to itself. So the set of geometric vectors with addition will determine the set of translations with a binary operation. What is the binary operation on translations corresponding to addition on geometric vectors?

### Exercise 2

Let the points  $O$ ,  $A$ ,  $B$ ,  $C$  and  $D$  in the  $xy$ -plane have co-ordinates  $(0, 0)$ ,  $(-2, 1)$ ,  $(3, 2)$ ,  $(2, 4)$  and  $(1, 5)$  respectively. Construct representatives of the following geometric vectors graphically.

- (i)  $\underline{OA} + \underline{OB}$
- (ii)  $\underline{OB} - \underline{BC}$
- (iii)  $\underline{OB} + \underline{BC} + \underline{CD} + \underline{DA} + \underline{AO}$



## Exercise 3

- (i) If the geometric vectors  $\underline{q}$  and  $\underline{h}$  are determined by the triples  $(1, -2, 4)$  and  $(2, 3, 1)$  respectively, show that  $\underline{q}$  and  $\underline{h}$  are perpendicular.
- (ii) Find another geometric vector  $\underline{\zeta}$  determined by the triple  $(x, y, z)$ , say, which is perpendicular to both  $\underline{q}$  and  $\underline{h}$ .
- (iii) Show that  $\underline{q}$ ,  $\underline{h}$  and  $\underline{\zeta}$  are linearly independent.

## 4.9 Answers to Exercises

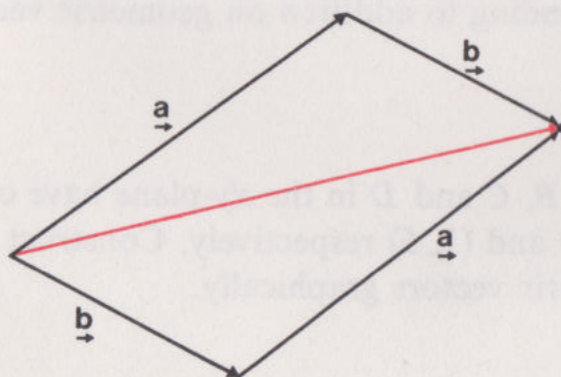
## Section 4.1

## Exercise 1

- (i) FALSE. Equal arrows must have the same length, direction and position.
- (ii) FALSE.  $\overrightarrow{AB}$  is an arrow, whereas  $\underline{AB}$  is a set of arrows.
- (iii) TRUE. Both the arrows  $\overrightarrow{FO}$  and  $\overrightarrow{ED}$  belong to the same geometric vector.
- (iv) FALSE. The arrows  $\overrightarrow{AO}$  and  $\overrightarrow{EF}$  (representatives of  $\underline{AO}$  and  $\underline{EF}$  respectively) are in opposite directions.
- (v) FALSE. The arrows  $\overrightarrow{BC}$  and  $\overrightarrow{AD}$  have different lengths.

## Section 4.2

## Exercise 1



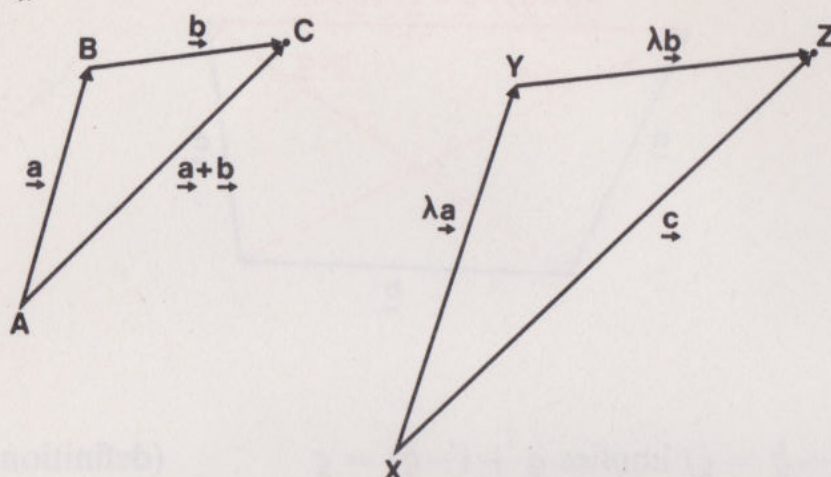
It is true that  $\underline{a} + \underline{b} = \underline{b} + \underline{a}$ , and therefore addition is commutative.







have the following diagram, where  $\underline{c} = \lambda \underline{a} + \lambda \underline{b}$ . We want to show that  $\underline{c} = \lambda(\underline{a} + \underline{b})$ .



Since  $\frac{XY}{AB} = \frac{YZ}{BC} = \lambda$ , and angle  $\hat{A}BC = \text{angle } X\hat{Y}Z$ , the two triangles are similar. Therefore

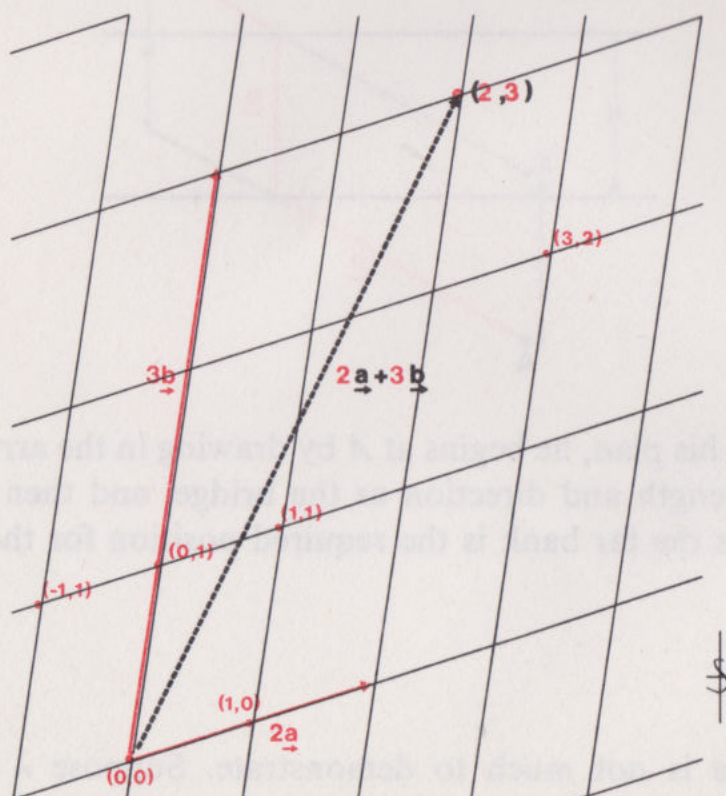
$$\frac{XZ}{AC} = \lambda$$

Further,  $XZ$  is parallel to  $AC$ , and therefore  $\underline{XZ} = \lambda \underline{AC}$ , i.e.

$$\underline{c} = \lambda(\underline{a} + \underline{b})$$

If  $\lambda < 0$ , we have a similar argument.

### Exercise 3





## Section 4.4

## Exercise 1

- (i)  $\vec{q} + (-\vec{q}) = \vec{0}$ , and hence  $\vec{q}$  and  $-\vec{q}$  are linearly dependent.  
 (ii) Choose the values  $\alpha = 0$  and  $\beta = 1$ , say, then

$$\alpha\vec{q} + \beta\vec{0} = 0\vec{q} + \vec{0} = \vec{0}$$

It follows that  $\vec{q}$  and  $\vec{0}$  are linearly dependent.

## Exercise 2

If  $\vec{q}$  and  $\vec{b}$  are linearly dependent, then we know that there are numbers  $\alpha$  and  $\beta$ , not both zero, such that

$$\alpha\vec{q} + \beta\vec{b} = \vec{0}$$

- (i) It follows that

$$\frac{\alpha}{3}(3\vec{q}) + \frac{\beta}{4}(4\vec{b}) = \vec{0}$$

and hence  $3\vec{q}$  and  $4\vec{b}$  are also linearly dependent.

- (ii)  $\vec{q} + \vec{b}$  and  $\vec{q} - \vec{b}$  are linearly dependent if we can find numbers  $\lambda$  and  $\mu$  not both zero, such that

$$\lambda(\vec{q} + \vec{b}) + \mu(\vec{q} - \vec{b}) = \vec{0}$$

i.e., such that

$$(\lambda + \mu)\vec{q} + (\lambda - \mu)\vec{b} = \vec{0}$$

Now  $\vec{q}$  and  $\vec{b}$  are linearly dependent so there are numbers  $\alpha$  and  $\beta$ , not both zero, such that

$$\alpha\vec{q} + \beta\vec{b} = \vec{0}$$

Suppose that we choose  $\lambda + \mu = \alpha$

$$\text{and } \lambda - \mu = \beta$$

$$\text{so that } \lambda = \frac{\alpha + \beta}{2} \quad \text{and} \quad \mu = \frac{\alpha - \beta}{2},$$

then if  $\lambda$  and  $\mu$  are not both zero we have shown that  $\vec{q} + \vec{b}$  and  $\vec{q} - \vec{b}$  are linearly dependent. But this follows at once, since  $\lambda = \mu = 0$  implies  $\alpha = \beta = 0$ , which we know to be false.



## Exercise 3

(i) Suppose that we can find numbers  $\alpha$  and  $\beta$  such that

$$\alpha(3\mathbf{a}) + \beta(4\mathbf{b}) = \mathbf{0}$$

i.e.

$$3\alpha(\mathbf{a}) + 4\beta(\mathbf{b}) = \mathbf{0}$$

But  $\mathbf{a}$  and  $\mathbf{b}$  are linearly independent, so  $3\alpha = 4\beta = 0$ , which implies that  $\alpha = \beta = 0$ . So  $3\mathbf{a}$  and  $4\mathbf{b}$  are linearly independent.

(ii) Yes. The proof is similar to part (i).

## Section 4.6

## Exercise 1

We have

$$\begin{aligned} \zeta \cdot (\mathbf{a} + \mathbf{b}) &= (\mathbf{a} + \mathbf{b}) \cdot \zeta & (\cdot \text{ is commutative}) \\ &= \mathbf{a} \cdot \zeta + \mathbf{b} \cdot \zeta & (\cdot \text{ is right-distributive over } +) \\ &= \zeta \cdot \mathbf{a} + \zeta \cdot \mathbf{b} & (\cdot \text{ is commutative}) \end{aligned}$$

## Section 4.8

## Exercise 1

The required binary operation is composition on the set of translations (i.e. perform one translation, and then the other). We can think of the set of geometric vectors being mapped on to the set of translations by a one-one function  $m$ :

$$m: \mathbf{a} \longmapsto f$$

This function  $m$  is then an *isomorphism* of the set of geometric vectors with addition to the set of translations with composition.

$$\begin{array}{ccc} (\mathbf{a}, \mathbf{b}) & \xrightarrow{+} & \mathbf{a} + \mathbf{b} \\ \downarrow m & & \downarrow m \\ (f, g) & \xrightarrow{\circ} & f \circ g = g \circ f \end{array}$$

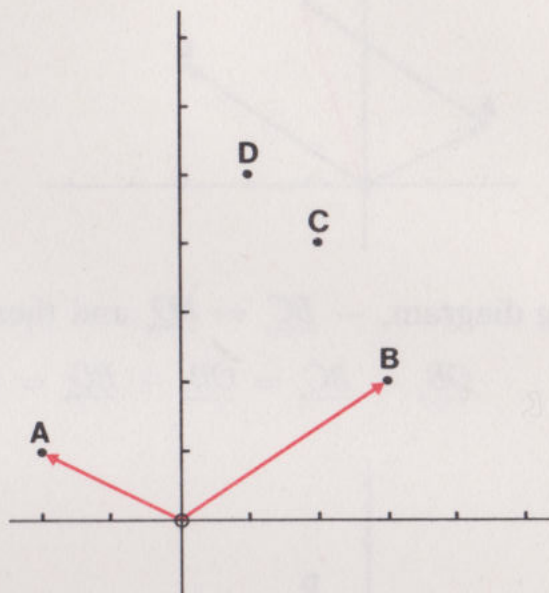
Notice that, although composition of function is not normally commutative, we know that it is commutative for the set of translations, because



it corresponds to the addition of geometric vectors, which is commutative.

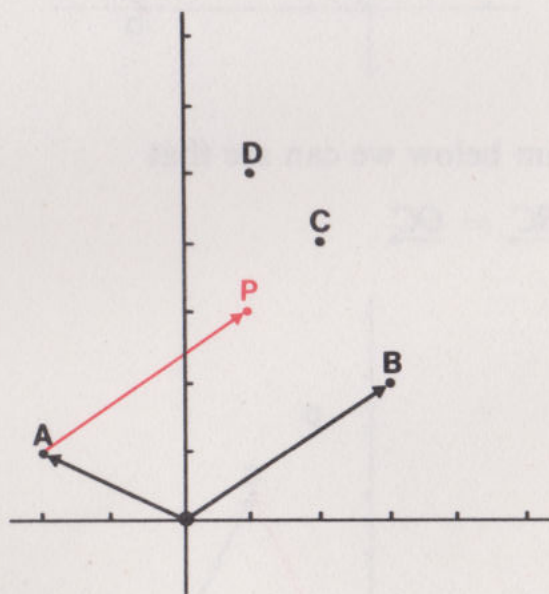
### Exercise 2

- (i)  $\overrightarrow{OA}$  and  $\overrightarrow{OB}$  are the two arrows shown below in red.



Our problem is to find  $\underline{OA} + \underline{OB}$ .

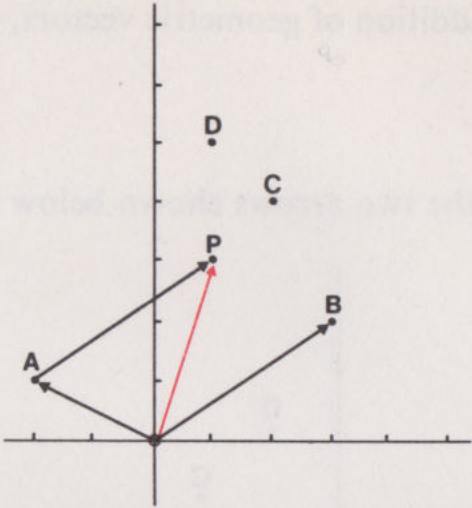
We can choose  $\overrightarrow{OA}$  as the representative of the geometric vector  $\underline{OA}$ ; then we need the arrow  $\overrightarrow{AP}$  as the representative of  $\underline{OB}$ .



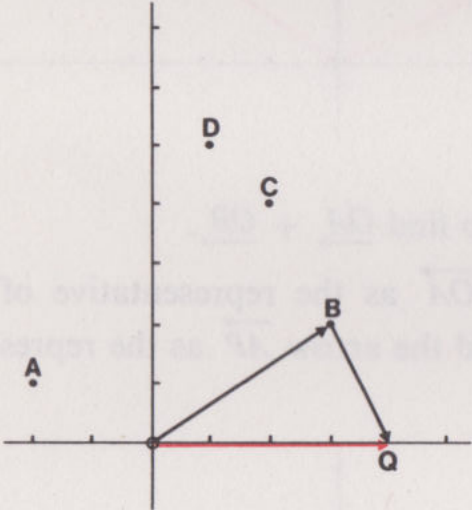
We can now see that

$$\underline{OA} + \underline{OB} = \underline{OA} + \underline{AP} = \underline{OP}$$

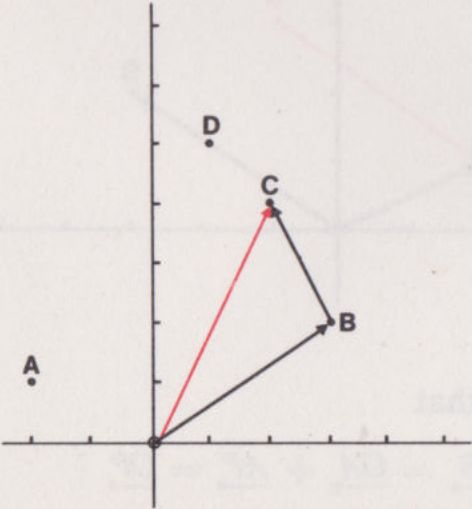




(ii) In the following diagram,  $-\underline{BC} = \underline{BQ}$  and therefore  
$$\underline{OB} - \underline{BC} = \underline{OB} + \underline{BQ} = \underline{OQ}$$

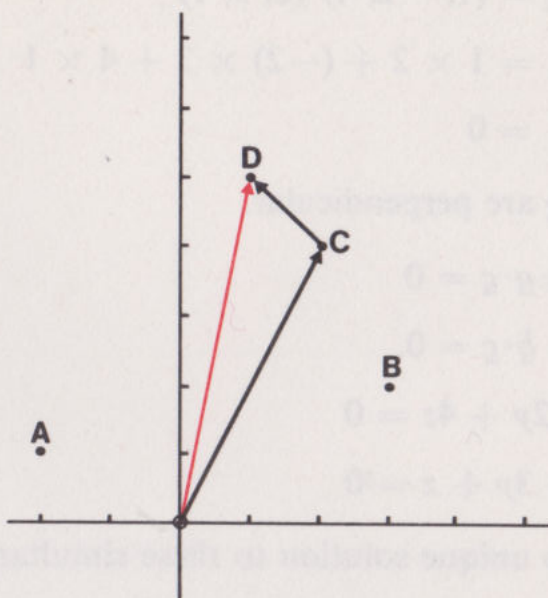


(iii) From the diagram below we can see that  
$$\underline{OB} + \underline{BC} = \underline{OC}$$

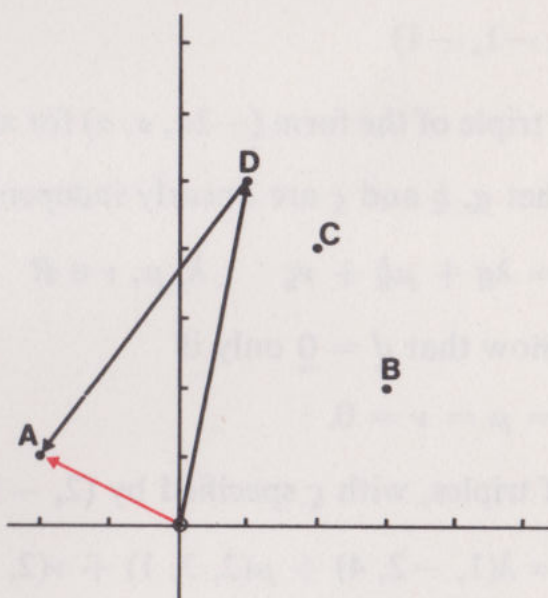




Then  $\underline{OC} + \underline{CD} = \underline{OD}$ ,



also  $\underline{OD} + \underline{DA} = \underline{OA}$ ,



but  $\underline{OA} + \underline{AO} = \underline{0}$ .

It follows that

$$\underline{OB} + \underline{BC} + \underline{CD} + \underline{DA} + \underline{AO} = \underline{0}$$

### Exercise 3

- (i) If  $\underline{a} \cdot \underline{b} = 0$ , then  $\underline{a}$  and  $\underline{b}$   
are perpendicular.



In this case

$$\begin{aligned}\underline{a} \cdot \underline{b} &= (1, -2, 4) \cdot (2, 3, 1) \\ &= 1 \times 2 + (-2) \times 3 + 4 \times 1 \\ &= 0\end{aligned}$$

i.e.  $\underline{a}$  and  $\underline{b}$  are perpendicular.

(ii) We require  $\underline{a} \cdot \underline{c} = 0$

and  $\underline{b} \cdot \underline{c} = 0$ .

So  $x - 2y + 4z = 0$

and  $2x + 3y + z = 0$

There is no unique solution to these simultaneous equations.

Examples of suitable triples are

$$(-2, 1, 1)$$

$$(2, -1, -1)$$

In fact any triple of the form  $(-2\alpha, \alpha, \alpha)$  for any  $\alpha \in R$  will be suitable.

(iii) To show that  $\underline{a}$ ,  $\underline{b}$  and  $\underline{c}$  are linearly independent, consider

$$\underline{d} = \lambda \underline{a} + \mu \underline{b} + \nu \underline{c} \quad \lambda, \mu, \nu \in R$$

We must show that  $\underline{d} = \underline{0}$  only if

$$\lambda = \mu = \nu = 0.$$

In terms of triples, with  $\underline{c}$  specified by  $(2, -1, -1)$

$$\begin{aligned}\underline{d} &= \lambda(1, -2, 4) + \mu(2, 3, 1) + \nu(2, -1, -1) \\ &= (\lambda + 2\mu + 2\nu, -2\lambda + 3\mu - \nu, 4\lambda + \mu - \nu)\end{aligned}$$

So we require  $\lambda + 2\mu + 2\nu = 0$

$$-2\lambda + 3\mu - \nu = 0$$

$$4\lambda + \mu - \nu = 0$$

Solving by elimination gives:

$$\lambda = \mu = \nu = 0$$

So  $\underline{a}$ ,  $\underline{b}$  and  $\underline{c}$  are linearly independent.



## CHAPTER 5 VECTOR SPACES

### 5.0 Introduction

In this chapter we define the mathematical structure called a *vector space*, of which geometric vectors form one example. We generalize some of the concepts mentioned in the previous chapter and then turn our attention to mappings of vector spaces, and, particularly those mappings of vector spaces which are morphisms. A particularly important subset of the domain of such morphisms is the set of elements which map to the zero vector in the codomain. We call this set of elements the *kernel* of the morphism, and it has some remarkable properties. We shall see that the kernel is itself a vector space and that (amongst other things) it provides us with a method of solving certain kinds of equations. If we know the kernel and one solution of such an equation, then we can easily generate every other solution.

This method has direct applications to the solution of problems in the theory of linear equations, in complex numbers, and in differential equations, and analogous methods can be developed in other areas of mathematics.

### 5.1 The Algebra of Lists

On various occasions in the previous chapter we saw that if we have a co-ordinate system, then there is a one-one correspondence between geometric vectors in two (or three) dimensions and ordered pairs (or triples) of numbers.

We developed some ways of combining geometric vectors, that is to say, we developed an algebra of geometric vectors, and we showed that there was a corresponding algebra for the ordered pairs or triples corresponding to the result. For example,

$$\begin{aligned}(a_1\mathbf{i} + a_2\mathbf{j} + a_3\mathbf{k}) + (b_1\mathbf{i} + b_2\mathbf{j} + b_3\mathbf{k}) \\ = (a_1 + b_1)\mathbf{i} + (a_2 + b_2)\mathbf{j} + (a_3 + b_3)\mathbf{k}\end{aligned}$$

we have

$$(a_1, a_2, a_3) + (b_1, b_2, b_3) = (a_1 + b_1, a_2 + b_2, a_3 + b_3).$$



We can regard these triples just as *ordered lists* and write them in any convenient form; for example, we can write

$$\begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} a_1 + b_1 \\ a_2 + b_2 \\ a_3 + b_3 \end{pmatrix} \quad \text{Equation (1)}$$

Such “vertical” lists are used later when we discuss matrices.

In the same way, the multiplication of a geometric vector by a scalar corresponds to

$$\lambda \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} \lambda a_1 \\ \lambda a_2 \\ \lambda a_3 \end{pmatrix}. \quad \text{Equation (2)}$$

So far these lists are just a way of specifying geometric vectors, but do they only give us an alternative notation, or do they suggest anything new? Let’s forget for a moment the origins of the lists. Equations (1) and (2) define ways of manipulating lists of numbers. There is no reason why we should always have only two or three elements in the list. Equations (1) and (2) can be extended to lists with more than three elements; for example, we can write

$$\begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_n \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{pmatrix} = \begin{pmatrix} a_1 + b_1 \\ a_2 + b_2 \\ a_3 + b_3 \\ \vdots \\ a_n + b_n \end{pmatrix}$$

and

$$\lambda \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} \lambda a_1 \\ \lambda a_2 \\ \lambda a_3 \\ \vdots \\ \lambda a_n \end{pmatrix}.$$



But this is rather futile if we only have a physical or mathematical interpretation when the lists contain not more than three elements. However, we can use these lists to describe situations other than the algebra of geometric vectors, and we can interpret results and concepts in one situation (for example, basis and linear independence) to give results and concepts in another. That is, we can establish *morphism* from one structure to another, and we shall go on to discuss the abstract structure which typifies all the exemplary situations.

What else can we represent by lists?

### Example 1 Polynomial Functions

Consider the set of all polynomial functions of the form

$$p: x \longmapsto ax^3 + bx^2 + cx + d \quad (x \in R)$$

where  $a, b, c$  and  $d$  are real numbers.

We can represent  $p$  by the four coefficients  $a, b, c$  and  $d$ , which we can arrange as a list:

$$\begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix}.$$

The addition of two such polynomial functions corresponds to the addition of the corresponding two lists. Thus if

$$p_1: x \longmapsto a_1x^3 + b_1x^2 + c_1x + d_1 \quad (x \in R)$$

and

$$p_2: x \longmapsto a_2x^3 + b_2x^2 + c_2x + d_2 \quad (x \in R)$$

then

$p_1 + p_2$  corresponds to the list

$$\begin{pmatrix} a_1 \\ b_1 \\ c_1 \\ d_1 \end{pmatrix} + \begin{pmatrix} a_2 \\ b_2 \\ c_2 \\ d_2 \end{pmatrix} = \begin{pmatrix} a_1 + a_2 \\ b_1 + b_2 \\ c_1 + c_2 \\ d_1 + d_2 \end{pmatrix}$$



and the function  $\lambda p$  corresponds to the list

$$\lambda \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} = \begin{pmatrix} \lambda a \\ \lambda b \\ \lambda c \\ \lambda d \end{pmatrix}$$

(Remember that  $a, b, c$  and  $d$  can be *any* real numbers, and so a function such as  $f: x \mapsto 0x^3 + 0x^2 + x + 1$  is included in this set of functions.) By considering polynomials of degree higher than three we would get examples of lists with more than four elements.

### Example 2 Finite Sequences

Consider the set of all finite sequences of real numbers with, say,  $n$  terms.

The methods of adding finite sequences and multiplying them by a number, which we met in Volume I, Chapter 6, can be seen to correspond to our algebra of lists, simply by rewriting a sequence

$$u_1, u_2, u_3, \dots, u_n$$

in the form of a list

$$\begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_n \end{pmatrix}$$

### Example 3 Solution of Differential Equations

In Volume 1, Chapter 8, we introduced the differentiation operator

$$D: f \mapsto f',$$

with domain the set of all real functions. We often write  $D^2f = f''$ , and similarly for higher derivatives. With this notation we are able to discuss more complicated operators such as

$$D^2 - 3D + 2,$$



which is defined by

$$(D^2 - 3D + 2):f \longmapsto f'' - 3f' + 2f.$$

Often in applied mathematics we are faced with the problem of finding a function which is mapped to a given function,  $g$  say, under such an operator. In other words, what function (or functions)  $f$  satisfy the equation

$$(D^2 - 3D + 2)f = g?$$

In terms of image values we require that

$$D^2f(x) - 3Df(x) + 2f(x) = g(x).$$

Equations of this kind are called *differential equations*. Suppose, for example, that  $g$  is the zero function, i.e.

$$g:x \longmapsto 0 \quad (x \in R)$$

then

$$f_1:x \longmapsto e^x \quad (x \in R)$$

is one function which satisfies this equation.

We know that

$$f_1''(x) = e^x \quad \text{and} \quad f_1'(x) = e^x$$

so that

$$(D^2 - 3D + 2)f_1(x) = e^x - 3e^x + 2e^x = 0.$$

Another function which satisfies the equation is  $f_2:x \longmapsto e^{2x} \quad (x \in R)$ . It can be shown that any solution of this differential equation has the form

$$\alpha f_1 + \beta f_2$$

where  $\alpha$  and  $\beta$  are real numbers, and  $f_1$  and  $f_2$  are the functions given above. If we take  $f_1$  and  $f_2$  as *basic solutions*, then we see that any solution

of the form  $\alpha f_1 + \beta f_2$  can be represented by the list  $\begin{pmatrix} \alpha \\ \beta \end{pmatrix}$ . The particular

solutions  $f_1$  and  $f_2$  can be represented by  $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$  and  $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$  respectively, and in

general the list  $\begin{pmatrix} \alpha \\ \beta \end{pmatrix}$  represents the function

$$x \longmapsto \alpha e^x + \beta e^{2x} \quad (x \in R).$$



You may like to verify that the lists

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} \gamma \\ \delta \end{pmatrix} = \begin{pmatrix} \alpha + \gamma \\ \beta + \delta \end{pmatrix}$$

and

$$\lambda \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \lambda\alpha \\ \lambda\beta \end{pmatrix}$$

also represent solutions of the equation.

#### Example 4 Interpolating Polynomials

The functions:

$$f_1: x \mapsto \frac{1}{2}x^2 - \frac{1}{2}x \quad (x \in R)$$

$$f_2: x \mapsto -x^2 + 1 \quad (x \in R)$$

$$f_3: x \mapsto \frac{1}{2}x^2 + \frac{1}{2}x \quad (x \in R)$$

have an interesting property. If we tabulate the images at  $-1$ ,  $0$  and  $1$ , we get the following table:

$x$	$-1$	$0$	$1$
$f_1(x)$	1	0	0
$f_2(x)$	0	1	0
$f_3(x)$	0	0	1

We can use these three functions to write down, without calculation, a formula for the quadratic function which has given values at  $-1$ ,  $0$  and  $1$ . For example, the quadratic function  $f$  which takes the values 2, 3 and 6 at  $-1$ ,  $0$  and  $1$  respectively, is given by

$$f: x \mapsto 2f_1(x) + 3f_2(x) + 6f_3(x) \quad (x \in R)$$

i.e.

$$f: x \mapsto x^2 + 2x + 3 \quad (x \in R)$$

We can use the idea of this example to derive Simpson's rule (see Volume 2, Chapter 4) very quickly. Suppose a curve passes through the three points  $(-1, g_{-1})$ ,  $(0, g_0)$ ,  $(1, g_1)$ . Then we know from this example that if we approximate to this curve by a parabola passing through these points, then that parabola is the graph of the function

$$g: x \mapsto g_{-1}f_1(x) + g_0f_2(x) + g_1f_3(x) \quad (x \in R)$$

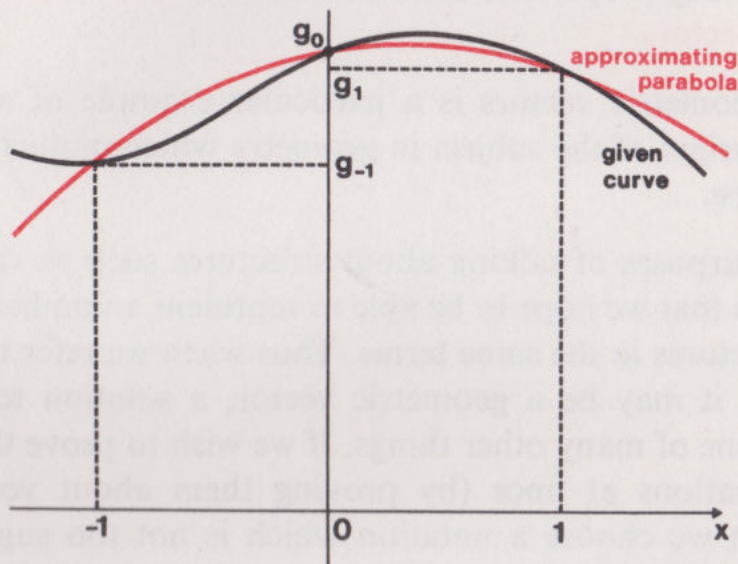


i.e.

$$g: x \mapsto g_{-1}(\tfrac{1}{2}x^2 - \tfrac{1}{2}x) + g_0(-x^2 + 1) + g_1(\tfrac{1}{2}x^2 + \tfrac{1}{2}x),$$

i.e.

$$g: x \mapsto x^2(\tfrac{1}{2}g_{-1} - g_0 + \tfrac{1}{2}g_1) + x(\tfrac{1}{2}g_1 - \tfrac{1}{2}g_{-1}) + g_0.$$



The area bounded by the parabola, the  $x$ -axis and the lines defined by  $x = -1$  and  $x = 1$  is

$$\begin{aligned} \int_{-1}^1 g &= \left[ x \mapsto \frac{x^3}{3} (\tfrac{1}{2}g_{-1} - g_0 + \tfrac{1}{2}g_1) + \frac{x^2}{2} (\tfrac{1}{2}g_1 - \tfrac{1}{2}g_{-1}) + g_0 x \right]_{-1}^1 \\ &= \tfrac{1}{3}(g_{-1} + 4g_0 + g_1) \end{aligned}$$

which is Simpson's rule with unit sub-interval length. (You may like to see if you can make the very slight modification to this piece of work which gives the more general form of Simpson's rule, where the strips are of width  $h$ .)

When a unifying idea emerges, it is often useful to strip it of all its original trappings and try to express the real essence of the idea. So we shall now extract the essential properties from our discussion so far.

## 5.2 Vector Spaces

We now generalize the discussion to an arbitrary set of elements, and construct a very general mathematical structure called a *vector space*.



A feature common to the geometric vectors and all the examples in the last section, is that, in each case, we had a set on which we could sensibly define *addition* and *multiplication by a scalar*.

We shall take the structure which we have developed on the set of geometric vectors as our model, and discuss an arbitrary set with operations called *addition* and *multiplication by a scalar* defined on it. If the structure has the following properties, then we call it a **vector space**, and we call its elements **vectors**.

The set of geometric vectors is a particular example of a vector space, and it is the origin of the subject in geometry which motivates this use of the word *space*.

One of the purposes of talking about structures such as vector spaces in the abstract is that we hope to be able to represent a number of apparently different structures in the same terms. Thus when we refer to a vector in a vector space, it may be a geometric vector, a solution to a differential equation or one of many other things. If we wish to prove theorems about all these situations at once (by proving them about vector spaces in general), then we choose a notation which is not too suggestive of any one example, and yet does not lose entirely the connection with our principal example, in this case geometric vectors. We therefore use underlined, lower case letters such as  $\underline{a}$  to represent a vector.

If  $\underline{v}, \underline{v}_1, \underline{v}_2, \underline{v}_3$  are any elements of a set  $V$ , and  $\alpha, \beta$  are any real numbers, we require the operations of *addition* of elements of  $V$  and *multiplication* of elements of  $V$  by a scalar to have the following properties (satisfy the following axioms):

- 1  $\underline{v}_1 + \underline{v}_2$  is a unique element of  $V$   
( $V$  is closed for addition)
- 2  $\underline{v}_1 + (\underline{v}_2 + \underline{v}_3) = (\underline{v}_1 + \underline{v}_2) + \underline{v}_3$   
(addition is associative)
- 3  $\underline{v}_1 + \underline{v}_2 = \underline{v}_2 + \underline{v}_1$   
(addition is commutative)
- 4 There is an element in  $V$ , which we call  $\underline{v}_0$ , such that
 
$$\underline{v} + \underline{v}_0 = \underline{v}$$
- 5  $\alpha \underline{v}$  is an element of  $V$
- 6  $\underline{v} + (-1)\underline{v} = \underline{v}_0$
- 7  $\alpha(\underline{v}_1 + \underline{v}_2) = (\alpha \underline{v}_1) + (\alpha \underline{v}_2)$
- 8  $(\alpha + \beta)\underline{v} = \alpha \underline{v} + \beta \underline{v}$
- 9  $(\alpha\beta)\underline{v} = \alpha(\beta \underline{v})$
- 10  $1 \times \underline{v} = \underline{v}$



These ten axioms are the axioms of a vector space. There are two important points to note. Strictly speaking we should call  $V$  a vector space over the real numbers or a real vector space, because vector spaces exist involving sets of scalars other than the set of real numbers; we shall discuss only vector spaces over the real numbers. Secondly, we have taken as implicit all the relevant properties of the real numbers, and these should really be stated along with the other axioms. Any other set with these properties can be taken as the set of scalars in place of the set of real numbers to give a different vector space.

The axioms of a vector space therefore consist of three sets of axioms:

- (i) those applying to the set of vectors only  
(1 to 4 above);
- (ii) those applying to the set of scalars only  
(not stated above: the missing axioms are the axioms of what is known in mathematics as a *field*);
- (iii) those which describe the interaction between the set of scalars (field) and the set of vectors  
(5 to 10 above).

We define an operation of subtraction of vectors by

$$\underline{v}_1 - \underline{v}_2 = \underline{v}_1 + (-1)\underline{v}_2.$$

From axiom 6, it follows that  $\underline{v}_1 - \underline{v}_1 = \underline{v}_0$ .

### The Zero Element

In a vector space, an element  $\underline{v}_0$  which satisfies axiom 4 is called a *zero element*. It follows from the axioms that in any vector space  $V$  there is only *one* zero element. For suppose there are two vectors  $\underline{v}_0$  and  $\underline{v}'_0$  which satisfy axiom 4. That is,

$$\underline{v} + \underline{v}_0 = \underline{v}$$

$$\underline{v} + \underline{v}'_0 = \underline{v},$$

where, in each equation,  $\underline{v}$  is *any* element of  $V$ . Let us put  $\underline{v} = \underline{v}'_0$  in the first equation, and  $\underline{v} = \underline{v}_0$  in the second equation. We obtain

$$\underline{v}'_0 + \underline{v}_0 = \underline{v}'_0$$

$$\underline{v}_0 + \underline{v}'_0 = \underline{v}_0.$$



By axiom 3,

$$\underline{v}'_0 + \underline{v}_0 = \underline{v}_0 + \underline{v}'_0$$

i.e.

$$\underline{v}_0 = \underline{v}'_0,$$

so the zero element is *unique*.

Since the zero element in a vector space behaves just like the zero geometric vector, we shall call this element the **zero vector** and denote it by  $\underline{0}$ , just as we had  $\underline{0}$  for the zero geometric vector. (Remember that, in terms of lists,  $\underline{0}$  is the list in which every entry is zero.)

Further properties of  $\underline{0}$  can be deduced from the axioms. For example, putting  $\underline{v}_2 = \underline{0}$  in axiom 7, we obtain

$$\alpha(\underline{v}_1 + \underline{0}) = (\alpha\underline{v}_1) + (\alpha\underline{0}).$$

By axiom 4,

$$\alpha\underline{v}_1 = \alpha\underline{v}_1 + \alpha\underline{0}$$

Now we add  $(-1)\alpha\underline{v}_1$  to both sides, and use axioms 2 and 3 to give

$$\alpha\underline{v}_1 + (-1)\alpha\underline{v}_1 = (\alpha\underline{v}_1 + (-1)\alpha\underline{v}_1) + \alpha\underline{0}$$

By axiom 6, we have

$$\underline{0} = \underline{0} + \alpha\underline{0}$$

Using axioms 3 and 4 and interchanging the sides of the equation, gives

$$\alpha\underline{0} = \underline{0},$$

where  $\alpha$  is any real number.

### Exercise 1

The set of all polynomial functions of degree  $n$  with the operations of addition of functions and multiplication of a function by a real number is not a vector space. Why not? Suggest a suitable modification to make it a vector space.

(HINT: A (real) polynomial function of degree  $n$  is a function of the form

$$x \longmapsto a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 \quad (x \in R)$$

in which the  $a_i$  are real numbers ( $i = 0, 1, 2, \dots, n$ ) and  $a_n \neq 0$ .)



## Exercise 2

In each of the following cases state whether the given set of lists forms a vector space for the operations of addition of lists and multiplication of a list by a scalar. In each case give reasons for your answer.

(i) The set of all lists  $\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$ , where  $x_1, x_2$  and  $x_3$  are positive real numbers.

(ii) The set of all lists  $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ , where  $x_1, x_2$  are real numbers and  $x_1 + x_2 = 0$ .

(iii) The set of all lists  $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ , where  $x_1$  and  $x_2$  are real numbers and  $x_1 < x_2$ .

(iv) The set of all lists  $\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$ , where  $x_1, x_2$  and  $x_3$  are real numbers such that

the function

$$f: t \longmapsto x_1 t^2 + x_2 t + x_3 \quad (t \in \mathbb{R})$$

satisfies  $f(k) = 0$ , where  $k$  is a fixed real number.

## Where Next?

In the case of geometric vectors, we introduced the idea of a *basis*. The development of this idea depended on the concepts of linear combination of vectors and linear dependence. We can extend these ideas to the more general concept of a vector space.

We also made passing reference to these ideas in the differential equation example: we shall see that every solution of

$$D^2 f(x) - 3Df(x) + 2f(x) = 0$$

can be represented in terms of two basic solutions, for example

$$f_1: x \longmapsto e^x \quad (x \in \mathbb{R})$$

$$f_2: x \longmapsto e^{2x} \quad (x \in \mathbb{R}).$$

This example raises a number of questions. How can we choose elements to use as a basis? How many elements do we need? If we can settle the question of how many elements we need—can we select that number of elements at random? Is there any test to see whether or not an arbitrarily chosen set of elements of the right number will do the job?



Before we extend our idea of a *basis* to a general vector space, we shall define linear dependence and independence in this context.

### Linear Dependence and Independence

The following definitions generalize the notion of linear dependence which we introduced for geometric vectors.

If  $\underline{v}_1, \underline{v}_2, \underline{v}_3, \dots, \underline{v}_n$  are vectors from any vector space, then an expression of the form

$$\alpha_1 \underline{v}_1 + \alpha_2 \underline{v}_2 + \alpha_3 \underline{v}_3 + \dots + \alpha_n \underline{v}_n,$$

where the  $\alpha$ 's are real numbers, is called a **linear combination** of vectors.

The set of vectors  $\{\underline{v}_1, \underline{v}_2, \dots, \underline{v}_n\}$  is said to be **linearly dependent** if and only if there exist real numbers  $\alpha_1, \alpha_2, \dots, \alpha_n$ , which are not all zero, such that

$$\alpha_1 \underline{v}_1 + \alpha_2 \underline{v}_2 + \alpha_3 \underline{v}_3 + \dots + \alpha_n \underline{v}_n = \underline{0}.$$

A set of vectors which is not linearly dependent is said to be *linearly independent*. We can define this term in a more positive way as follows.

A set of vectors  $\{\underline{v}_1, \underline{v}_2, \underline{v}_3, \dots, \underline{v}_n\}$  is **linearly independent** if and only if

$$\alpha_1 \underline{v}_1 + \alpha_2 \underline{v}_2 + \alpha_3 \underline{v}_3 + \dots + \alpha_n \underline{v}_n = \underline{0}$$

$$\text{implies } \alpha_1 = \alpha_2 = \alpha_3 = \dots = \alpha_n = 0.$$

Remember that we use the terms *dependent* and *independent* in this way because we can express some members of a linearly dependent set in terms of the others. For example, if  $\alpha_1$  is not zero, we can use the axioms of a vector space to write

$$\alpha_1 \underline{v}_1 + \alpha_2 \underline{v}_2 + \alpha_3 \underline{v}_3 + \dots + \alpha_n \underline{v}_n = \underline{0}$$

in the form

$$\alpha_1 \underline{v}_1 = (-\alpha_2) \underline{v}_2 + (-\alpha_3) \underline{v}_3 + \dots + (-\alpha_n) \underline{v}_n,$$

and then divide by  $\alpha_1$  to give:

$$\underline{v}_1 = \frac{-\alpha_2}{\alpha_1} \underline{v}_2 + \frac{-\alpha_3}{\alpha_1} \underline{v}_3 + \dots + \frac{-\alpha_n}{\alpha_1} \underline{v}_n,$$

i.e.  $\underline{v}_1$  depends on the other vectors. In general, if a set of vectors is linearly dependent, *some* of the vectors in the set (not necessarily every vector, because *some* of the  $\alpha$ 's may be zero) can be expressed in terms of the others. In other words, some of the elements in the set are redundant.



## Exercise 3

In each of the following parts a set of vectors is given. In each case state whether or not the set is linearly independent. In those cases where the set is linearly dependent, express one of the vectors in the set as a linear combination of the others.

$$(i) \quad \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix},$$

with the usual operations of addition and multiplication by a scalar

for lists. The zero vector in this case is  $\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$ .

(ii) The functions

$$f: x \longmapsto x \quad (x \in R),$$

$$g: x \longmapsto x^2 \quad (x \in R),$$

with the operations of addition of functions and multiplication of a function by a real number. The zero vector in this case is

$$\underline{0}: x \longmapsto 0 \quad (x \in R).$$

$$(iii) \quad \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 3 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 0 \\ 5 \end{pmatrix}$$

with the usual operations for combining "lists".

## Exercise 4

If the set of vectors  $\underline{v}_1, \underline{v}_2, \underline{v}_3, \dots, \underline{v}_n$  is linearly independent, show that if

$$\alpha_1 \underline{v}_1 + \alpha_2 \underline{v}_2 + \alpha_3 \underline{v}_3 + \dots + \alpha_n \underline{v}_n = \beta_1 \underline{v}_1 + \beta_2 \underline{v}_2 + \dots + \beta_n \underline{v}_n$$

then

$$\alpha_1 = \beta_1, \quad \alpha_2 = \beta_2, \quad \dots, \quad \alpha_n = \beta_n.$$

Notice that this result implies that a vector  $\underline{v}$  cannot be expressed in two *different* ways as a linear combination of a set of linearly independent vectors.



### 5.3 Bases and Dimension of a Vector Space

In section 4.4 we saw that it is possible to select two geometric vectors in a plane, and then to specify every geometric vector in the plane as a linear combination of those two. Similarly, in three dimensions we need to select three geometric vectors. We called such a set a *basis*, and we now wish to extend the same idea to a vector space.

The set of vectors  $\{\underline{v}_1, \underline{v}_2, \dots, \underline{v}_m\}$  is said to **span** the vector space  $V$  if for each element  $\underline{w}$  in  $V$  we can find scalars  $\alpha_1, \alpha_2, \dots, \alpha_m$ , such that

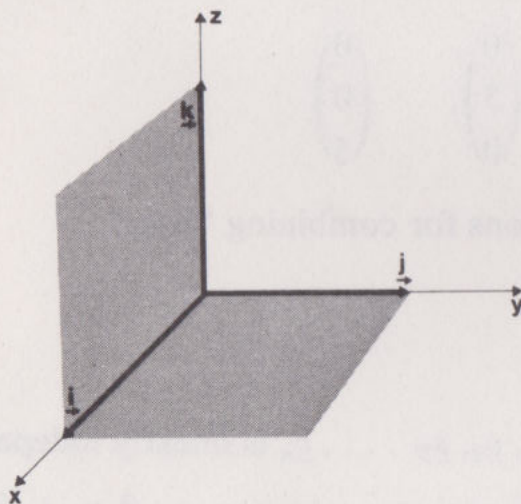
$$\underline{w} = \alpha_1 \underline{v}_1 + \alpha_2 \underline{v}_2 + \alpha_3 \underline{v}_3 + \dots + \alpha_m \underline{v}_m.$$

If the set of vectors  $\{\underline{v}_1, \underline{v}_2, \dots, \underline{v}_n\}$  is linearly independent and spans the vector space  $V$ , then we say that it forms a **basis** for  $V$ , and  $\underline{v}_1, \underline{v}_2, \dots, \underline{v}_n$  are called **base vectors**.

Essentially a basis contains the minimum number of elements which are required to span the space. In Exercise 5.2.4 we saw that any vector can be expressed in a *unique* way as a linear combination of the elements of a basis.

For example, the set  $\{\underline{i}, \underline{j}, \underline{k}\}$  spans the three-dimensional geometric vector space, because each geometric vector  $\underline{r}$  can be expressed in the form

$$\underline{r} = x\underline{i} + y\underline{j} + z\underline{k}.$$



Here  $\underline{i}, \underline{j}$  and  $\underline{k}$  play the parts of  $\underline{v}_1, \underline{v}_2$  and  $\underline{v}_3$ , and we know that it is possible to find the appropriate values  $x, y$  and  $z$  which play the parts of  $\alpha_1, \alpha_2$  and  $\alpha_3$ . Any set of geometric vectors containing  $\underline{i}, \underline{j}$  and  $\underline{k}$  and other geometric vector(s) would also span the space, but it would not form a basis, since such a set would be linearly dependent (the other geometric vector(s) would be redundant).



*Exercise 1*

Show that the set  $\left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}$  is a basis for the set of all triples of real numbers.

As a result of this last exercise we have two distinct bases, namely  $\left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}$  and  $\left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}$  for the same vector space, the space of all triples, and in this case both bases consist of three vectors. In fact, although we shall not prove it here, this always happens: for any two sets of base vectors for the same vector space, there are always the same number of vectors in each basis. This enables us to make the following definition.

If  $\{\underline{v}_1, \underline{v}_2, \dots, \underline{v}_n\}$  is a basis for a vector space  $V$ , then we say that the vector space is of **dimension**  $n$ .

If it is impossible to find a finite number of elements of a vector space  $V$  which form a basis for  $V$ , and  $V \neq \{\underline{0}\}$ , then we say that  $V$  has *infinite dimensions*.

It is in fact also true that *any* set of  $n$  linearly independent vectors in a vector space  $V$  of dimension  $n$  is a basis for  $V$ .

If we assume these results, then we can see that, since  $\left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\}$  is a basis of the vector space of ordered pairs of real numbers, this vector space is therefore of dimension 2. The set  $\left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}$  is a basis for the set of ordered triples, and this vector space is therefore of dimension 3. Let us look now at some non-geometric examples.

*Example 1*

The set of all polynomial functions of degree 2 or less, i.e. of the form:

$$f: x \longmapsto ax^2 + bx + c \quad (x \in R)$$



where  $a, b, c \in R$ , forms a vector space with the operations of addition of functions and multiplication of a function by a real number.

We can find many sets of three vectors in this vector space which are linearly independent. One such set, which is particularly simple, consists of the vectors

$$\underline{f}_1: x \longmapsto 1 \quad (x \in R),$$

$$\underline{f}_2: x \longmapsto x \quad (x \in R),$$

$$\underline{f}_3: x \longmapsto x^2 \quad (x \in R).$$

Any other quadratic function can be expressed in terms of these three, and hence they form a basis for the vector space. The dimension of the space is therefore 3. The function

$$\underline{f}: x \longmapsto 3x^2 - 2x + 4 \quad (x \in R)$$

can be written as a linear combination of the base elements:

$$\underline{f} = 3\underline{f}_3 - 2\underline{f}_2 + 4\underline{f}_1$$

(We have underlined the  $\underline{f}$ 's because we want to emphasize the fact that we are considering the functions to be elements of a vector space.)

### Example 2

We stated in Example 5.1.3 that any solutions of the equation

$$D^2f(x) - 3Df(x) + 2f(x) = 0$$

can be expressed in terms of the two solutions

$$\underline{f}_1: x \longmapsto e^x \quad (x \in R),$$

$$\underline{f}_2: x \longmapsto e^{2x} \quad (x \in R).$$

In other words, these two functions span the space of solutions. Since the two solutions,  $\underline{f}_1$  and  $\underline{f}_2$  are independent, the set of all solutions of the equation forms a vector space of dimension 2. (If you are worried about the statement that  $\underline{f}_1$  and  $\underline{f}_2$  are independent, you might like to try to prove it.)

## 5.4 Mapping One Vector Space to Another

In Chapter 1 we began by mapping sets one to another, and we introduced the concept of a *function*. With this concept we found that we could think about problems which are more sophisticated and have wider application than simple arithmetic calculations. It is the same with vector



spaces; the topic becomes richer and more interesting when we introduce mappings from one vector space to another.

Some mappings of vector spaces to vector spaces are simply equivalent to a change of notation.

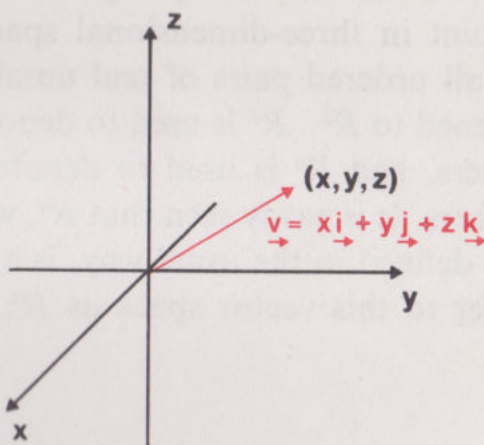
### Example 1

We know that the set of geometric vectors forms a vector space, as does the set of all lists with three elements.

If  $\underline{v} = x\underline{i} + y\underline{j} + z\underline{k}$ , then the mapping

$$n:\underline{v} \mapsto \begin{pmatrix} x \\ y \\ z \end{pmatrix},$$

with domain the set of geometric vectors, simply gives us an alternative notation.



( $\underline{i}$ ,  $\underline{j}$  and  $\underline{k}$  are the unit geometric vectors in the directions of the  $x$ ,  $y$  and  $z$  Cartesian axes respectively: see section 5.3).

Under this mapping we have

$$n:\underline{i} \mapsto \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

$$n:\underline{j} \mapsto \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

$$n:\underline{k} \mapsto \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$



We can calculate  $n(\underline{v})$  for any geometric vector  $\underline{v}$  once we know the images of the base vectors.

Notice that if we want a mapping to define a new notation for the elements in its domain then the mapping must be one-one.

Another “notational” mapping is the mapping of ordered lists to ordered pairs defined by

$$\begin{pmatrix} x \\ y \end{pmatrix} \longmapsto (x, y),$$

or ordered lists to ordered triples defined by

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \longmapsto (x, y, z),$$

which simply states the obvious fact that, given a co-ordinate system, we can represent a two-element list by a point in a plane and a three-element list by a point in three-dimensional space. We already have a name for the set of all ordered pairs of real numbers; we call it  $R \times R$ . This is usually shortened to  $R^2$ .  $R^3$  is used to denote the set of all ordered triples of real numbers, and  $R^n$  is used to denote the set of all ordered  $n$ -tuples of real numbers. It is easily seen that  $R^n$ , with addition and multiplication by a scalar defined in the usual way, is a vector space of dimension  $n$ . We shall refer to this vector space as  $R^n$ , leaving the operations to be understood.

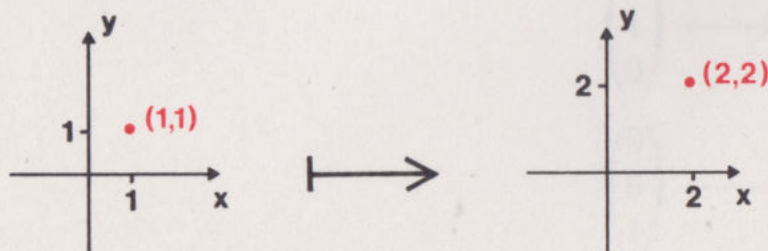
### Example 2

Consider the mapping of  $R^2$  to  $R^2$  defined by

$$f:(x, y) \longmapsto (2x, 2y).$$

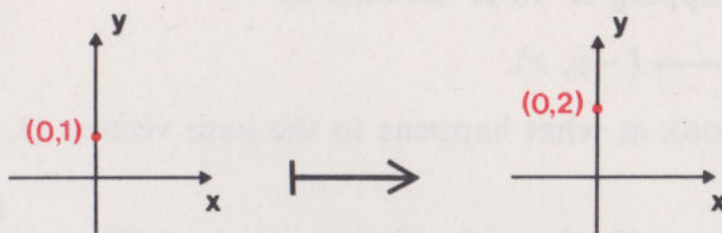
What effect does this mapping have? One way to start to answer this question is to consider what happens to a few particular elements.

Take the element  $(1, 1)$ ; then  $f(1, 1) = (2, 2)$ .





For the element  $(0, 1)$ , we have  $f(0, 1) = (0, 2)$ .



In general, we see that the point  $P$  is mapped on to the point  $P'$  in the same direction away from the origin  $O$ , where  $OP' = 2OP$ . (Does any point remain unchanged?)

Consider a set of elements in the plane, such as

$$\{(x, y): x^2 + y^2 = 1\}.$$

These points lie on a circle of unit radius, and therefore their images under this mapping will lie on a circle of radius 2. We can verify this algebraically as follows. Suppose  $(x, y) \mapsto (u, v)$ , then  $u = 2x$  and  $v = 2y$ , and so if  $x$  and  $y$  satisfy the equation

$$x^2 + y^2 = 1,$$

$u$  and  $v$  must satisfy the equation

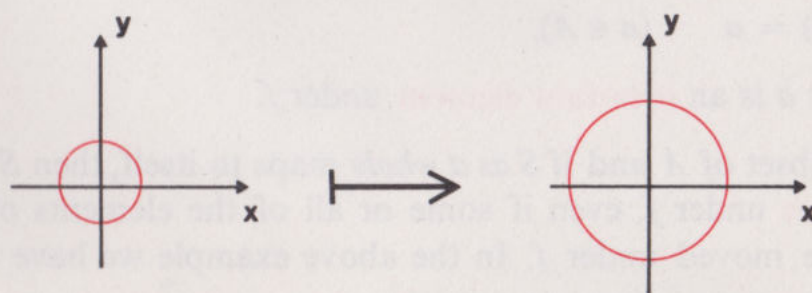
$$\left(\frac{u}{2}\right)^2 + \left(\frac{v}{2}\right)^2 = 1$$

i.e.  $u^2 + v^2 = 4$

Thus

$$\begin{aligned} \{(x, y): x^2 + y^2 = 1\} &\mapsto \{(u, v): u^2 + v^2 = 4\} \\ &= \{(x, y): x^2 + y^2 = 4\} \end{aligned}$$

(We have re-written this set in terms of  $x$  and  $y$  so that we can draw the following diagram.)





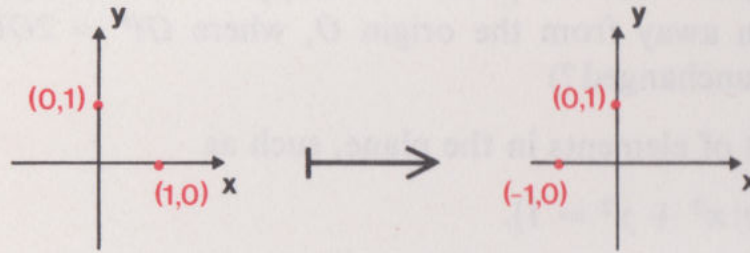
**Example 3**

Consider the mapping  $R^2$  to  $R^2$  defined by

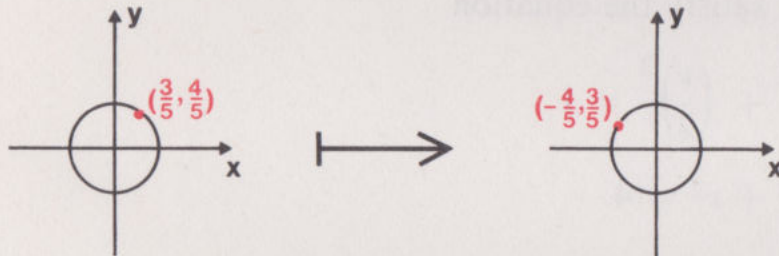
$$(x, y) \longmapsto (-y, x).$$

Let us have a look at what happens to the base vectors  $(1, 0)$  and  $(0, 1)$ . We have

$$(1, 0) \longmapsto (0, 1) \quad \text{and} \quad (0, 1) \longmapsto (-1, 0).$$



The mapping has the effect of rotating the base vectors through an angle  $\frac{\pi}{2}$  counter-clockwise about the origin, and this is indeed the effect on the entire plane.



In this example, any circle centred at the origin maps on to itself. Every point of the set  $\{(x, y) : x^2 + y^2 = 1\}$  moves (for example,  $(\frac{3}{5}, \frac{4}{5}) \longmapsto (-\frac{4}{5}, \frac{3}{5})$ ), but the set itself remains unchanged.

Example 3 provides a particular example of a simple but important mathematical concept. Given a function  $f$  of a set  $A$  to itself, then if, for instance,

$$f(a) = a \quad (a \in A),$$

we say that  $a$  is an **invariant element** under  $f$ .

If  $S$  is a subset of  $A$  and if  $S$  as a whole maps to itself, then  $S$  is called an **invariant set** under  $f$ , even if some or all of the elements of the subset  $S$  of  $A$  are moved under  $f$ . In the above example we have an invariant



circle, i.e. an invariant set of points. The concept of invariance can be even more general. For example, under a translation of the plane, the distance between two points is invariant; that is, if  $P$  and  $Q$  map to  $P'$  and  $Q'$  respectively, then the length of  $PQ = \text{length of } P'Q'$ .

### Exercise 1

Under the mapping

$$f: (x, y) \longmapsto (2x, 2y),$$

in Example 2, we have seen that circles centred at the origin are not invariant. But some lines are invariant. Which lines?

### Example 4

Consider the mapping of  $R^2$  to  $R^2$  defined by

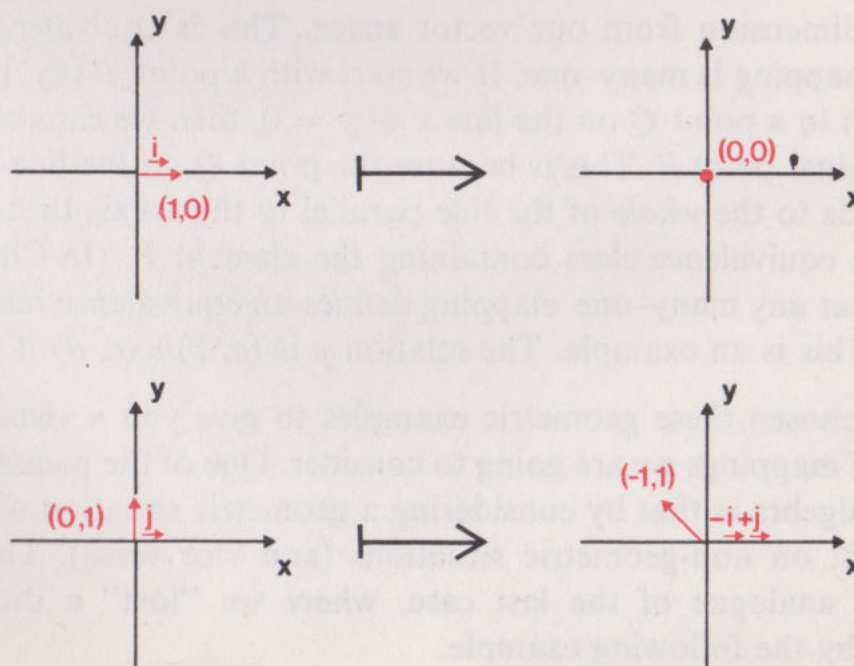
$$(x, y) \longmapsto (-y, x).$$

From the previous two examples, it seems that, to see the effect of the mapping, it is probably worth seeing what happens to a set of base vectors. Again, to simplify the arithmetic, we choose the base vectors  $(1, 0)$  and  $(0, 1)$ . We find that

$$(1, 0) \longmapsto (0, 0)$$

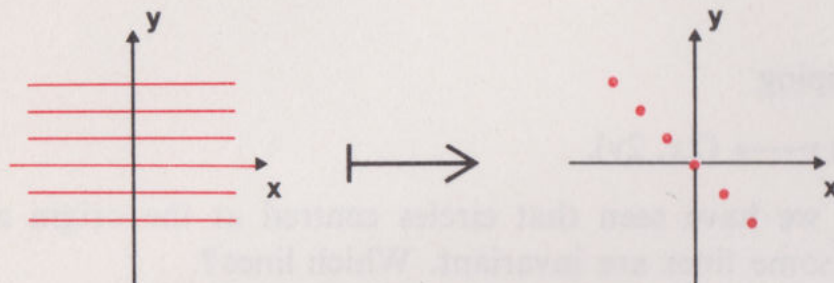
$$(0, 1) \longmapsto (-1, 1),$$

so that in terms of the corresponding geometric vectors  $\underline{i}$  and  $\underline{j}$ , we have that  $\underline{i}$  maps to the zero vector, but  $\underline{j}$  maps to  $-\underline{i} + \underline{j}$ .





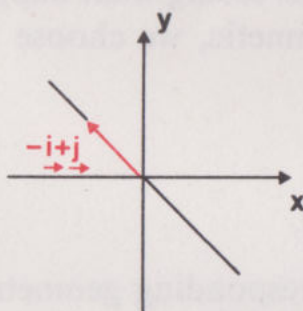
We shall look at this mapping in a little more detail. What happens to the  $x$ -axis? On this axis  $y$  is zero and so every point on the  $x$ -axis maps to  $(0, 0)$ : the entire  $x$ -axis shrinks into the origin. What about the line  $y = 1$ ? Every point on this line maps to the point  $(-1, 1)$ .



In fact, the entire plane maps on to the line whose equation is

$$x + y = 0$$

In terms of geometric vectors, every element in the image set can be represented in terms of the vector  $-i + j$ .



The image set has dimension 1, and so the effect of the mapping is to “lose” a dimension from our vector space. This is equivalent to saying that this mapping is many-one. If we start with a point,  $P$  say, in the plane and map it to a point  $Q$  on the line  $x + y = 0$ , then we cannot map back to the original point  $P$ . This is because the point  $Q$  on the line  $x + y = 0$  corresponds to the *whole* of the line parallel to the  $x$ -axis through  $P$ ; this line is the equivalence class containing the element  $P$ . (In Chapter 2 we showed that any many-one mapping defines an equivalence relation on its domain. This is an example. The relation  $\rho$  is  $(a, b) \rho (c, d)$  if  $b = d$ .)

We have chosen these geometric examples to give you a visualization of the sort of mappings we are going to consider. One of the pleasant features of linear algebra is that by considering a geometric situation we can often throw light on non-geometric situations (and vice versa). Thus, a non-geometric analogue of the last case, where we “lost” a dimension, is provided by the following example.



## Example 5

Let  $P_3$  be the vector space of all polynomial functions of degree 3 or less. (This space has dimension 4, and a set of base vectors is given below.) The differentiation operator:

$$D: p \longmapsto Dp \quad (p \in P_3)$$

maps the space  $P_3$  to the space  $P_2$  (which has dimension 3). In this case we could take as a basis for  $P_3$  the set of functions:

$$\left. \begin{array}{ll} \underline{f}_0: x \longmapsto 1 & (x \in R) \\ \underline{f}_1: x \longmapsto x & (x \in R) \\ \underline{f}_2: x \longmapsto x^2 & (x \in R) \\ \underline{f}_3: x \longmapsto x^3 & (x \in R). \end{array} \right\} \text{Basis for } P_3$$

This set of functions maps to the set

$$\left. \begin{array}{ll} \underline{f}'_0: x \longmapsto 0 & (x \in R) \\ \underline{f}'_1: x \longmapsto 1 & (x \in R) \\ \underline{f}'_2: x \longmapsto 2x & (x \in R) \\ \underline{f}'_3: x \longmapsto 3x^2 & (x \in R). \end{array} \right\} \text{Basis for } P_2$$

Like the previous example, this mapping is many-one. (For example, all the functions of the form:

$$\underline{f}: x \longmapsto x^2 + a \quad (x \in R),$$

where  $a \in R$ , map to the function  $\underline{f}'_2$ .)

Note that  $\underline{f}'_0$  is the *zero vector* in  $P_2$ . It cannot belong to any basis for  $P_2$ , since a basis must be a linearly independent set of three vectors. For consider the set  $\{\underline{f}'_0, \underline{g}, \underline{h}\}$ , where  $\underline{g}, \underline{h} \in P_2$ . Since

$$\alpha \underline{f}'_0 + 0\underline{g} + 0\underline{h} = \underline{f}'_0,$$

where  $\alpha$  is any *non-zero* real number, we see that any set of three elements containing  $\underline{f}'_0$  is linearly *dependent*.

## Exercise 2

We seem to be pinning a lot of faith in choosing a convenient basis. Is the choice of basis important? We shall resolve this difficulty later, but one point can be considered here.



In Example 5, instead of  $f_0$  and  $f_1$ , we could choose  $g_0$  and  $g_1$ , where

$$g_0: x \longmapsto 1 + x \quad (x \in R)$$

$$g_1: x \longmapsto 1 - x \quad (x \in R),$$

and then none of the base vectors maps to the zero vector under  $D$ . Is  $\{g'_0, g'_1, f'_3\}$  a basis for  $P_2$ ?

### Exercise 3

Fill in the gaps in the solution to the following problem.

#### Problem

If  $T$  is a mapping from  $R^2$  to  $R^2$ , and

$$T: (x, y) \longmapsto \left(\frac{x}{2}, \frac{y}{3}\right),$$

describe the image of the ellipse

$$\{(x, y): 9x^2 + 4y^2 = 36\}.$$

#### Solution of Problem

Put  $u = \frac{x}{2}$  and  $v = \frac{y}{3}$

Since  $x$  and  $y$  satisfy the equation

$$9x^2 + 4y^2 = 36$$

$u$  and  $v$  satisfy the equation

$$\boxed{\phantom{9u^2 + 4v^2}} = 36 \quad (\text{i})$$

i.e.

$$\boxed{\phantom{u^2 + v^2}} = 1 \quad (\text{ii})$$

i.e.

$$\{(x, y): 9x^2 + 4y^2 = 36\} \longmapsto \{(u, v): \boxed{\phantom{u^2 + v^2}}\} \quad (\text{iii})$$

The image set corresponds to a circle centred at

$$\boxed{\phantom{(0, 0)}} \text{ and with radius } \boxed{\phantom{1}} \quad (\text{iv})$$



## 5.5 Morphisms

There are many interesting and useful results concerning mappings of vector spaces, but the most fruitful field of study consists of those mappings which are *homomorphisms* or *isomorphisms* (and therefore, of necessity, *functions*).

If  $V$  is a vector space, the two operations we have defined on the elements of  $V$  are addition of vectors and multiplication of a vector by a scalar. Suppose a function  $T$  maps a vector space  $V$ , with an addition operation  $+_v$ , to a vector space  $U$ , with an addition operation  $+_u$ . The additive structure will be preserved if, for any vectors  $\underline{v}_1$  and  $\underline{v}_2$  in  $V$ ,

$$T(\underline{v}_1 +_v \underline{v}_2) = T(\underline{v}_1) +_u T(\underline{v}_2)$$

Since we have been abusing the symbol  $+$  constantly (we have defined all sorts of methods of addition and called them all  $+$ ), we shall continue to do so, and we drop the suffices  $u$  and  $v$  from the addition symbols. We then have

$$T(\underline{v}_1 + \underline{v}_2) = T(\underline{v}_1) + T(\underline{v}_2) \quad \text{Equation (1)}$$

as the condition that  $T$  should be a morphism for the addition operations.

For the other operation we require that

$$T(\alpha \underline{v}) = \alpha T(\underline{v}), \quad \text{Equation (2)}$$

for any real number  $\alpha$  and any vector  $\underline{v} \in V$ .

Equations (1) and (2) are the condition that  $T$  should be a morphism from the vector space  $V$  to the vector space  $U$ .\*

The two equations can be combined to give the following equation:

$$T(\alpha_1 \underline{v}_1 + \alpha_2 \underline{v}_2) = \alpha_1 T(\underline{v}_1) + \alpha_2 T(\underline{v}_2), \quad \text{Equation (3)}$$

for any real numbers  $\alpha_1, \alpha_2$ , and any vectors  $\underline{v}_1$  and  $\underline{v}_2 \in V$ .

In many books, a mapping of a vector space to a vector space is called a *transformation*, and when the mapping is a morphism it is called a **linear transformation**. This is another example of calling a particular type of mapping by a special name (cf. the word *operator* used in Volume 1).

\* Previously we said that a morphism  $f$  was a mapping from  $A$  to  $f(A)$ . Here,  $T(V)$  may be a proper subset of  $U$ . (In fact, if  $T$  is a morphism,  $T(V)$  is a vector space, but we have not proved this yet.)



*Exercise 1*

Which of the following mappings are morphisms? (Take the operations in the various vector spaces to be the usual ones.)

- (i) The mapping of  $R^2$  to  $R^2$  such that

$$T:(x_1, x_2) \longmapsto (x_2, x_1)$$

- (ii) The mapping of  $R^2$  to  $R^2$  such that

$$T:(x_1, x_2) \longmapsto (x_1^2, x_2^2)$$

- (iii) The mapping of the set of all geometric vectors in a plane to  $R$  such that

$$T:\underline{x} \longmapsto q \cdot \underline{x},$$

where  $q$  is a given geometric vector, and the dot stands for the inner product as defined in Chapter 4.

*Exercise 2*

Let  $L$  be a morphism from a vector space  $V$  to a vector space  $U$ .

Complete the gaps in the proof of the following theorem.

## THEOREM

If the zero element of  $V$  is  $\underline{v}_0$  (i.e.  $\underline{v}_0$  is the element for which  $\underline{v} + \underline{v}_0 = \underline{v}$  for any  $\underline{v} \in V$ ), and if  $\underline{u}_0$  is the zero element of  $U$ , then  $L(\underline{v}_0) = \underline{u}_0$ .

## PROOF

Since  $\underline{v} + \underline{v}_0 = \underline{v}$

$$L(\underline{v} + \underline{v}_0) = L\left(\boxed{\phantom{v}}\right) \tag{a}$$

But  $L$  is a morphism, so

$$L(\underline{v} + \underline{v}_0) = L\left(\boxed{\phantom{v}}\right) + L\left(\boxed{\phantom{v}}\right) \tag{b}$$

From (a) and (b),  $L(\underline{v}) + L(\underline{v}_0) = L(\underline{v})$ ,

so subtracting  $L(\underline{v})$  from both sides, we see that

$$L(\underline{v}_0) \text{ is the } \boxed{\phantom{u}} \text{ element of } U, \tag{c}$$



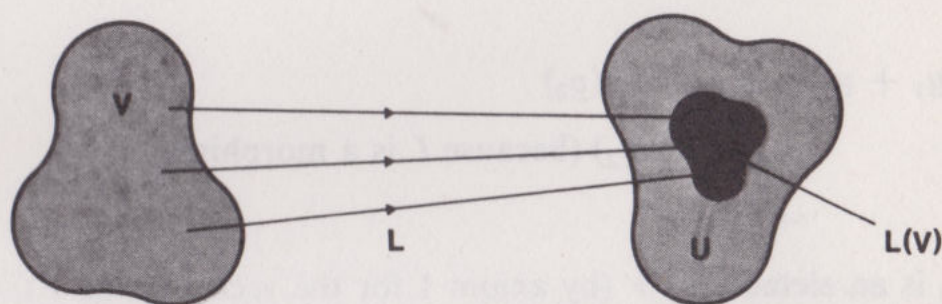
Confirm this result for each of the mappings which are morphisms in Exercise 1.

Exercise 2 shows that under a morphism the zero element in the domain vector space is mapped to the zero element in the codomain vector space. We shall use this result in proving the following theorem, which is fundamental to the study of morphisms of vector spaces.

#### THEOREM

If  $L$  is a morphism from a vector space  $V$  to a vector space  $U$ , then  $L(V)$  is a subset of  $U$  which is *itself* a vector space.

( $L(V)$  may be  $U$  or a proper subset of  $U$ .)



#### METHOD OF PROOF

We have to prove that the set  $L(V)$ , with the operations of the vector space  $U$ , satisfies the vector space axioms listed on page 158.

Axioms 2, 3, 6, 7, 8, 9 and 10 are statements about all elements of a vector space, and since  $U$  is a vector space we do not have to check these axioms for  $L(V)$ . On the other hand, axiom 4 is a statement that a particular kind of element (the zero element) belongs to a vector space. Clearly, the zero element of  $U$  will not belong to *every* subset of  $U$ , so we have to prove that it belongs to the particular subset  $L(V)$ . Axioms 1 and 5 concern *closure*. If  $L(V)$  is to be a vector space, then any combination of elements in  $L(V)$  must give resulting elements still in  $L(V)$ . This again is not necessarily true for *any* subset of  $U$ , so we must check it for  $L(V)$ .

#### PROOF

We have to prove that axioms 1, 4 and 5 hold for  $L(V)$ . We have three pieces of information:

- (i)  $V$  is a vector space;
- (ii)  $U$  is a vector space;
- (iii)  $L$  is a morphism.



We have used (ii) to dispose of axioms 2, 3, 6, 7, 8, 9 and 10, but we have not yet used (i) and (iii).

We have proved that axiom 4 holds for  $L(V)$ : in Exercise 2 we proved that under a morphism the zero element  $\underline{v}_0$  in the domain  $V$  maps to the zero element  $\underline{u}_0$  in the codomain  $U$ . So the zero element of  $U$  belongs to  $L(V)$ .

Let us have a look at axiom 1; we must show that  $L(V)$  is closed under addition. If  $\underline{u}_1$  and  $\underline{u}_2$  are any elements of  $L(V)$ , then there are elements  $\underline{v}_1$  and  $\underline{v}_2$  in  $V$  such that

$$\underline{u}_1 = L(\underline{v}_1),$$

$$\underline{u}_2 = L(\underline{v}_2).$$

Then

$$\begin{aligned}\underline{u}_1 + \underline{u}_2 &= L(\underline{v}_1) + L(\underline{v}_2) \\ &= L(\underline{v}_1 + \underline{v}_2) \text{ (because } L \text{ is a morphism)} \\ &= L(\underline{v}_3),\end{aligned}$$

where  $\underline{v}_3$  is an element of  $V$  (by axiom 1 for the vector space  $V$ );  $L(\underline{v}_3)$  is an element of  $L(V)$ , so  $\underline{u}_1 + \underline{u}_2$  belongs to  $L(V)$ , and  $L(V)$  is closed.

The other closure axiom, number 5, is easily checked.

If

$$\underline{u} = L(\underline{v}),$$

then

$$\begin{aligned}\alpha \underline{u} &= \alpha L(\underline{v}) \\ &= L(\alpha \underline{v}) \text{ (because } L \text{ is a morphism).}\end{aligned}$$

Therefore  $\alpha \underline{u} \in L(V)$ .

This completes the proof.

If a subset of a vector space  $U$  is itself a vector space, then we call it a **vector subspace** of  $U$ .

### Exercise 3

(i) The mapping from  $R^2$  to  $R^2$  defined by

$$L:(x_1, x_2) \longmapsto (-x_2, x_1)$$

is a morphism. Prove directly by verifying the axioms that  $L(R^2)$  is a vector space.



(ii) The mapping from  $R^2$  to  $R^2$  defined by

$$T:(x_1, x_2) \longmapsto (x_1^2, x_2)$$

is not a morphism. Show that  $T(R^2)$  is not a vector space by finding an axiom which is not satisfied.

(Note that we have *not* proved that if  $T:V \longrightarrow U$  is *not* a morphism, then  $T(V)$  is *not* a vector space. For a general mapping  $T$ , we do not know anything about  $T(V)$  if  $T$  is not a morphism.)

### Summary

So far we have considered mappings of one vector space to another, and we have concentrated our attention on those mappings which are morphisms. A morphism has the property that the image set itself is a vector space: we have seen that the properties of commutativity, associativity and the zero element are carried over from the domain to its image set by a morphism.

An interesting feature of *some* of the morphisms we have met is that they map a vector space on to an image set which has a lower dimension. For example, we have had mappings of planes to lines, polynomials of degree  $n$  or less to polynomials of degree  $n - 1$  or less, and so on. Two questions arise. What has happened to the “lost” dimensions? Can we predict in advance when we are going to “lose” a dimension? We shall look at these questions in the next section.

## 5.6 The Kernel

Let us have another look at Example 5.4.4 in which we saw that the morphism

$$L:(x, y) \longmapsto (-y, y) \quad ((x, y) \in R^2)$$

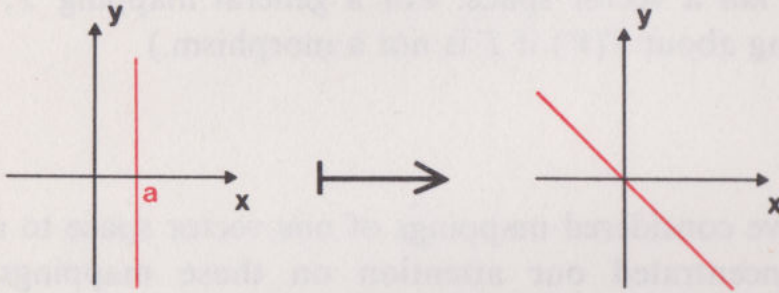
maps the plane to a line. First, we looked at a particular basis, and saw that one of the base vectors mapped to the zero element  $(0, 0)$  in the codomain. We then investigated the mapping by looking to see what happened to particular subsets of the plane. We saw that any line parallel to the  $x$ -axis mapped to a single point.

This raises two questions. First, is it significant that we lose *one* basis vector and we lose *one* dimension? Secondly, it is all very well in this simple case to pick out a few significant subsets that tell us such a lot.



We picked them out because we knew their properties. Consider now the images of the lines parallel to the  $y$ -axis in the domain. *Any* such line maps to the *entire* image set, for suppose we take the line for which  $x = a$ , then

$$L:(a, y) \longmapsto (-y, y) \quad (y \in \mathbb{R}).$$



By considering certain subsets of the domain, we find that we can obtain information about  $L$ . Is there any particular subset which we can most profitably consider? That is, can we describe a subset in the domain which will give us information about  $L$  in a form which we can interpret easily? If so, can we extract any general feature which will help us with other examples?

The clue is in our observation about the loss of a basis vector. The vector  $(1, 0)$  maps to  $(0, 0)$ , but of course it is not only this vector which “shrinks” to zero—every multiple of  $(1, 0)$  maps to  $(0, 0)$ . (Because the mapping is a morphism,  $L(\alpha(1, 0)) = \alpha L(1, 0) = (0, 0)$ .) So a whole set maps to  $(0, 0)$ . Why consider this particular set? Try the next exercise.

### Exercise 1

The vectors  $(0, 1)$  and  $(2, 2)$  form a basis for  $\mathbb{R}^2$ . Calculate  $L(0, 1)$  and  $L(2, 2)$ , where  $L$  is the mapping we have been discussing:

$$L:(x, y) \longmapsto (-y, y) \quad ((x, y) \in \mathbb{R}^2).$$

What happens to the linear independence of  $(0, 1)$  and  $(2, 2)$ ?

Exercise 1 shows us that, in our original basis, the choice of a vector which mapped to  $(0, 0)$  was purely fortuitous. (See page 171.) It may so happen, as in this exercise, that none of the base vectors maps to  $(0, 0)$ , even though we “lose” a dimension. But here the whole set  $\{(\alpha, 0) : \alpha \in \mathbb{R}\}$  maps to  $(0, 0)$ , whether or not one of its elements is in the basis.

It seems, then, that the set which maps to  $(0, 0)$  tells us something about



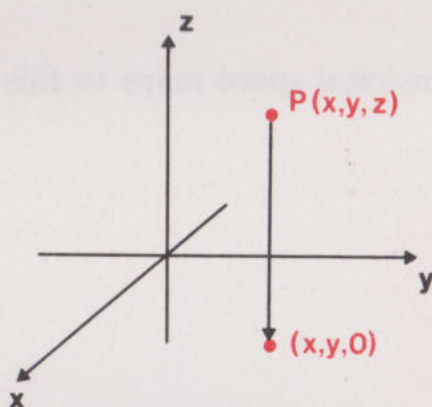
the “lost” dimension. In this case the set which maps to  $(0, 0)$  has dimension 1 (every element of the set can be obtained as a scalar multiple of  $(1, 0)$ ), and we lose just one dimension. Let us have a look at two more examples, one where we again lose one dimension and one where we lose more than one dimension. We shall again take geometrical examples because geometrical situations are easy to visualize.

### Example 1

The mapping

$$L:(x, y, z) \longmapsto (x, y, 0)$$

is a morphism of  $R^3$  to  $R^3$ . The image of any point  $P$  is the point at the foot of the perpendicular from  $P$  to the plane with equation  $z = 0$ . Thus the domain maps to a plane.



This morphism maps a 3-dimensional space to a 2-dimensional space: we lost one dimension. The set which maps to the zero element  $(0, 0, 0)$  in the codomain is the set  $\{(0, 0, z): z \in R\}$ , that is, the  $z$ -axis. This set is itself a vector space and its dimension is one, the same number as the number of “lost” dimensions.

### Example 2

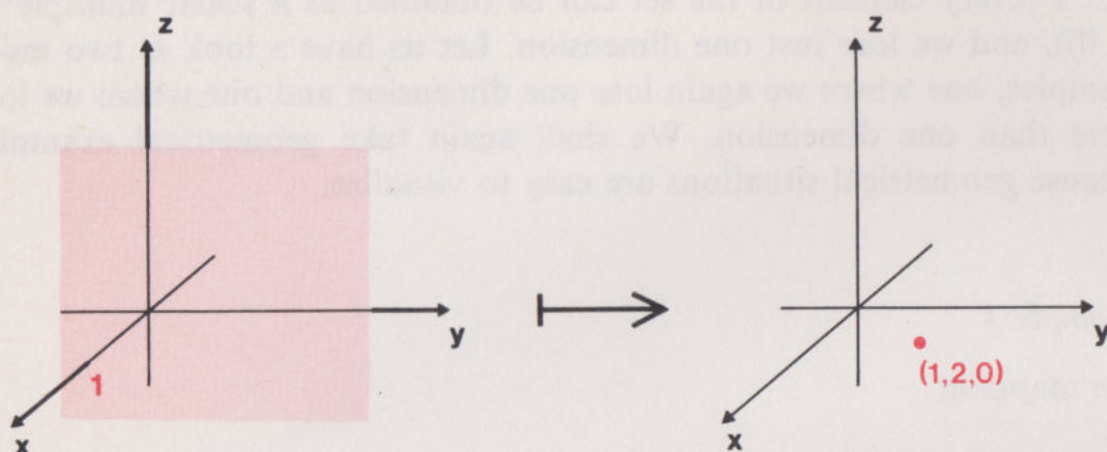
The mapping

$$L:(x, y, z) \longmapsto (x, 2x, 0)$$

is a morphism of  $R^3$  to  $R^3$ . The image of the point  $(x, y, z)$  depends only on its  $x$ -co-ordinate. Thus  $(1, 2, 3)$ ,  $(1, 6, 7)$ ,  $(1, 6, 99)$  all map to the point



$(1, 2, 0)$ . Every point in the plane with equation  $x = 1$  maps to this point.

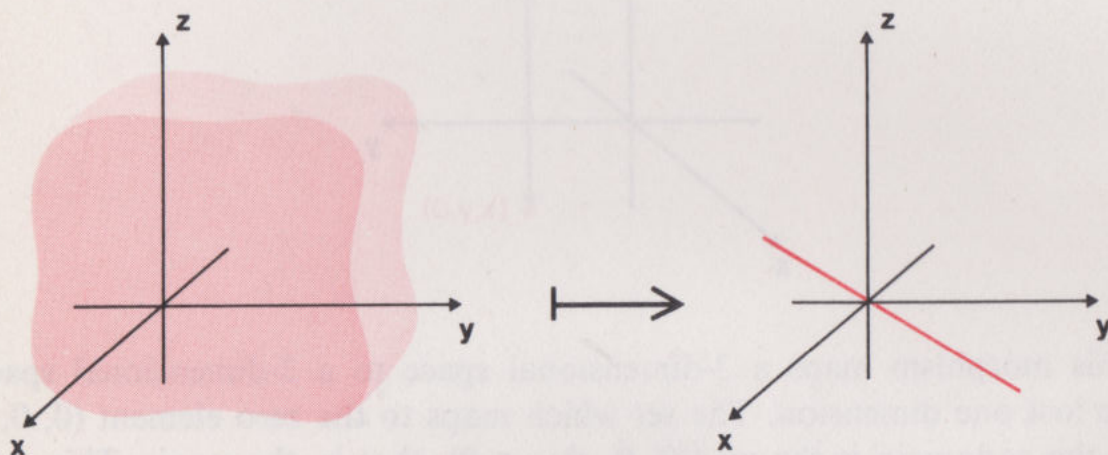


Similarly, every point on the plane with equation  $x = 2$  maps to the point  $(2, 4, 0)$ , and so on. Every plane perpendicular to the  $x$ -axis maps to a point on the line defined by the equations:

$$2x - y = 0$$

$$z = 0,$$

and the entire three-dimensional space maps to this complete line.



Thus

$$L(R^3) = \{(x, y, z) : 2x - y = 0, z = 0\}$$

The three-dimensional domain maps to a space of dimension 1. We seem to have lost two dimensions.

Which set maps to the zero element? In this case the zero element is  $(0, 0, 0)$ , and the set which maps to  $(0, 0, 0)$  is the set  $\{(x, y, z) : x = 0\}$ , i.e. the  $yz$ -plane; this is itself a vector space and has dimension two. Notice that in this simple case we can use the “basis argument” again—if we can find an appropriate basis. Taking the set of vectors  $\{(1, 0, 0)$ ,



$(0, 1, 0), (0, 0, 1)\}$  as a basis, we see that both  $(0, 1, 0)$  and  $(0, 0, 1)$  map to  $(0, 0, 0)$ , and so we “lose” two base vectors. In fact we “lose” any vector which can be expressed as a linear combination of these two vectors (i.e. the points in the plane with equation  $x = 0$ ), because

$$\begin{aligned} L(\alpha(0, 1, 0) + \beta(0, 0, 1)) \\ = \alpha L(0, 1, 0) + \beta L(0, 0, 1) = (0, 0, 0), \end{aligned}$$

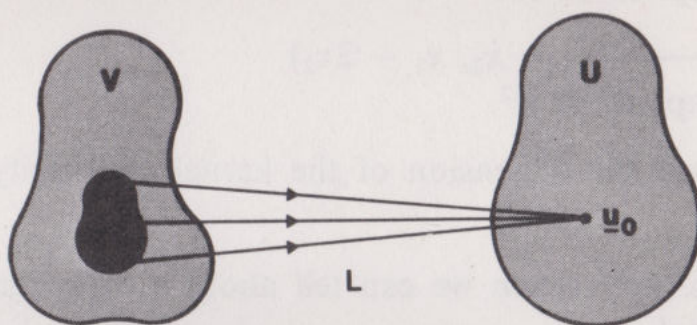
since  $L$  is a morphism.

We have seen that *the subset of the domain which maps to the zero element in the codomain* plays an important part, so we now give it a name.

If  $L$  is a morphism of a vector space  $V$  to a vector space  $U$ , and if  $\underline{u}_0$  is the zero element in  $U$ , then the set

$$\{\underline{v} : \underline{v} \in V, L(\underline{v}) = \underline{u}_0\}$$

is called the **kernel**\* of the morphism. (Another name which is in common use for this set is **the null space**.) We shall denote the kernel by the letter  $K$ .



There is one important point to notice here. We defined a *basis* of a vector space  $V$  to be a linearly independent set of vectors in  $V$  which spans  $V$ . We defined the *dimension* of  $V$  to be the number of elements in a basis. Now the kernel of a morphism may only contain the zero element of  $V$ ; i.e. we may have  $K = \{\underline{0}\}$ .

We write  $\underline{0}$  instead of  $\underline{v}_0$  here, because you may like to refer to our discussion of the vector space  $\{\underline{0}\}$  in section 5.2 (page 160). There, we proved that

$$\alpha \underline{0} = \underline{0},$$

where  $\alpha$  is any real number. This means that  $\{\underline{0}\}$  is a *linearly dependent set*; that is,  $\{\underline{0}\}$  *does not possess a basis*. We adopt the following definition.

\* This is in accordance with the ordinary use of *kernel* for the *nucleus* or *core* of a structure.



The dimension of the zero vector space,  $\{0\}$ , is zero.

The kernel has some quite remarkable properties. We have already hinted at two of them which are printed in red below.

(1) The kernel itself is a vector space.

We have shown that  $L(V)$  is itself a vector space, and in Examples 1 and 2 we have seen that

(2)  $(\text{dimension of } L(V)) = (\text{dimension of } V) - (\text{dimension of kernel})$

The first of these results is not difficult to prove: we have set it as one of the additional exercises for this chapter. The second result, which is called the *Dimension Theorem* for vector spaces, can be proved with the techniques at our disposal, but we feel it more appropriate to leave the proof out of an introductory text of this nature.

### Exercise 2

Find the kernel of each of the following morphisms:

- (i)  $T:(x_1, x_2, x_3) \longmapsto (x_1, x_2, 0)$   
where  $T$  maps  $R^3$  to  $R^3$ .
- (ii)  $T:(x_1, x_2) \longmapsto (x_1 + x_2, x_1 - 2x_2)$   
where  $T$  maps  $R^2$  to  $R^2$ .

In each case find the dimension of the kernel and verify statement (2) above.

It is remarkable how much we can tell about a morphism just by considering its kernel.

Let  $L$  be a morphism of a vector space  $V$  to a vector space  $U$ , and let  $K$  be its kernel. If  $\underline{k} \in K$  and  $\underline{v} \in V$ , then

$$\begin{aligned} L(\underline{v} + \underline{k}) &= L(\underline{v}) + L(\underline{k}) && (L \text{ is a morphism}) \\ &= L(\underline{v}) + \underline{u}_0 && (\text{definition of } K) \\ &= L(\underline{v}) && (\text{axiom 4 for } U), \end{aligned}$$

where  $\underline{u}_0$  is the zero element in  $U$ . So  $\underline{v}$  and  $\underline{v} + \underline{k}$ , where  $\underline{k}$  is any element of the kernel, have the same image.

Suppose now that we want to find *all* the elements in  $V$  which map to a given element  $\underline{u} \in U$ , and that we know one such element  $\underline{v}$ , i.e.

$$L(\underline{v}) = \underline{u}$$

Then we know immediately that  $\underline{v} + \underline{k}$ , for all  $\underline{k} \in K$ , are such elements,



and the remarkable thing is that they are in fact *all* the elements which map to  $\underline{u}$ . We can prove this as follows. Suppose  $\underline{v}_1 \in V$  maps to  $\underline{u}$ , i.e.  $L(\underline{v}_1) = \underline{u}$ . Then consider  $\underline{v}_1 + (-1)\underline{v}$ . We have

$$\begin{aligned} L(\underline{v}_1 + (-1)\underline{v}) &= L(\underline{v}_1) + (-1)L(\underline{v}) && (L \text{ is a morphism}) \\ &= \underline{u} + (-1)\underline{u} && (\text{by hypothesis}) \\ &= \underline{u}_0 && (\text{axiom 6 for } U). \end{aligned}$$

But the kernel  $K$  contains all those elements which map to  $\underline{u}_0$ , so

$$\underline{v}_1 + (-1)\underline{v} = \underline{k}_1$$

for some  $\underline{k}_1 \in K$ . By axiom 1 for  $V$ ,  $\underline{k}_1$  is *unique*. Adding  $\underline{v}$  to both sides, we get

$$\underline{v}_1 = \underline{v} + \underline{k}_1,$$

and so  $\underline{v}_1$  is of the form  $\underline{v} +$  some element of the kernel. This result has important consequences: for instance, if the kernel contains an infinite number of elements, then we know that an infinite number of elements of  $V$  map to *any* given element of  $L(V)$ . Furthermore, if the kernel contains just one element (which will necessarily be the zero element in  $V$ ), then we know immediately that  $L$  is one-one, i.e. an isomorphism. The following example applies this discussion.

### Example 3

We apply the above ideas to finding the solution of the equations

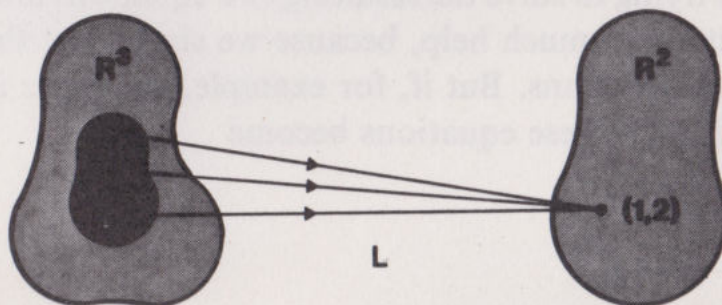
$$2x + 3y - z = 1$$

$$x + y - z = 2.$$

One way of expressing the problem is to say that we want to find the set of triples  $(x, y, z)$  which satisfy these equations. If  $L$  is the mapping from  $R^3$  to  $R^2$  defined by

$$L:(x, y, z) \longmapsto (2x + 3y - z, x + y - z).$$

then we want to find the set which maps to  $(1, 2)$ .





We have seen that we can describe the set which maps to any particular element when we know the kernel and one element of the set. In the context of this example, this means that, if we can find one solution to the equations, we shall be able to find *all* the solutions, simply by adding to that solution each element of the kernel. So we have to find *one solution* and we have to find the *kernel*.

### *One Solution*

If we give  $x$  or  $y$  or  $z$  a particular value, then the equations will be reduced to two equations in two unknowns, which we can solve easily.

For example, if we put  $z = 0$ , then we obtain the two equations

$$2x + 3y = 1$$

$$x + y = 2,$$

which we can solve to give

$$x = 5, \quad y = -3.$$

So one solution of the original equations

$$2x + 3y - z = 1$$

$$x + y - z = 2$$

is

$$x = 5, y = -3, z = 0.$$

### *The Kernel*

The kernel,  $K$ , is the set of triples  $(x, y, z)$  which map to  $(0, 0)$ , i.e. which satisfy the equations

$$2x + 3y - z = 0$$

$$x + y - z = 0$$

Just as before, we can solve these equations by giving  $x$  or  $y$  or  $z$  a particular value and then trying to solve the resulting two equations in two unknowns; but this time it is not much help, because we simply get the *one* solution, and we want *all* solutions. But if, for example, we give  $z$  a general value and put  $z = k$ , then these equations become

$$2x + 3y = k$$

$$x + y = k,$$

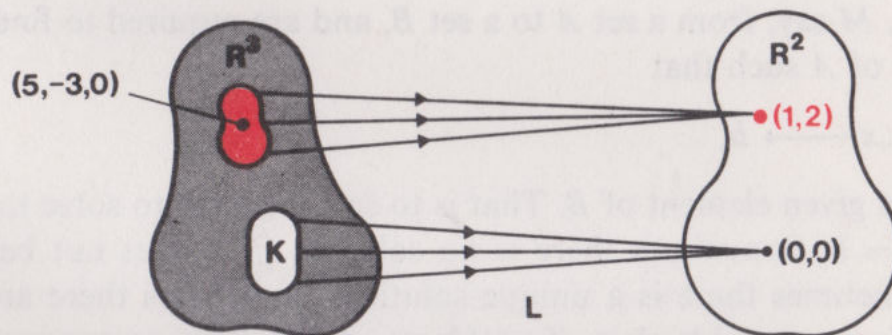


which we can solve to give

$$x = 2k, \quad y = -k.$$

So one element of the kernel is  $(2k, -k, k)$ , and by varying  $k$  we get all the elements. So  $K$  is the set

$$\{(2k, -k, k) : k \in R\}.$$



We want to find the red shaded set in the above diagram.

We know one element—how do we find the others?

Any solution of the original equations

$$2x + 3y - z = 1$$

$$x + y - z = 2$$

is obtained by adding an element of the kernel to  $(5, -3, 0)$ .

So the complete solution set is

$$\{(5 + 2k, -3 - k, k) : k \in R\},$$

and the theory we have developed assures us that this set contains *all* the possible solutions to the original equations.

(Check that these *are* solutions by substituting into the original equations.)

Notice that we can solve related problems like

$$2x + 3y - z = 7$$

$$x + y - z = 99,$$

where the right-hand sides of the equations are changed, very quickly—all we need to do is to find one particular solution and then add on each of the elements of the (same) kernel. We shall discuss problems of this type in considerable detail later.



*Exercise 3*

By putting  $x = 0$ , find a particular solution of the equations

$$2x + 3y - z = 2$$

$$x + y + z = 1$$

Find the solution set of the equations.

Problems such as the following often arise in mathematics. We are given a mapping,  $M$  say, from a set  $A$  to a set  $B$ , and are required to find all the elements  $x$  of  $A$  such that

$$M: x \longmapsto b,$$

where  $b$  is a given element of  $B$ . That is to say, we have to solve the equation  $M(x) = b$ . Sometimes there is no solution (if  $b$  does not belong to  $M(A)$ ); sometimes there is a unique solution; sometimes there are many solutions. In a very wide class of problems,  $A$  and  $B$  are vector spaces and  $M$  is a morphism. We have just seen that in these cases we can (in principle) adopt a standard procedure. We first find the kernel,  $K$ , that is, the solution set of

$$M(\underline{x}) = \underline{0}.$$

(This is usually a much easier problem than the original one.) We then find any one particular solution of

$$M(\underline{x}) = \underline{b},$$

and then combine this with each element of  $K$ , to get the complete solution set.

**Summary**

We have concentrated our attention on those mappings of vector spaces which are morphisms. Under a morphism the image of a vector space is itself a vector space. A particularly important subset of the domain of such morphisms is the set of elements which map to the zero vector in the codomain. We call this set of elements the *kernel* of the morphism, and it has some remarkable properties. It is itself a vector space; its dimension tells us the difference between the dimensions of the domain and the image space; it tells us whether the morphism is many-one or one-one; it provides us with a method of solving equations such as  $L(\underline{v}) = \underline{u}$ , where  $L$  is a morphism of a vector space. If  $L$  is a homomorphism then it is, by definition, a many-one mapping, and so we can expect this equation to have more than one solution.



We saw that if we know the kernel of the morphism and if we know one solution of the equation  $L(\underline{v}) = \underline{u}$ , then we can easily generate every other solution by using the elements of the kernel.

These results are listed below.

Let  $L$  be a morphism from a vector space  $V$  to a vector space  $U$ . Then,

- (i)  $L(V)$  is a vector subspace of  $U$ ;
- (ii)  $K = \{\underline{k} \in V : L(\underline{k}) = \underline{u}_0\}$  is a vector subspace of  $V$ , called the *kernel* of the morphism;
- (iii)  $(\text{dimension of } V) = (\text{dimension of } L(V)) + (\text{dimension of } K)$ ;
- (iv) if  $\underline{v}_1$  is a solution of  $L(\underline{v}) = (\underline{u})$ , then the set of all solutions of this equation is  $\{\underline{v}_1 + \underline{k} : \underline{k} \in K\}$ .

## 5.7 Additional Exercises

### Exercise 1

If  $V$  is a vector space with zero vector  $\underline{0}$ , show that  $\{\underline{0}\}$  is also a vector space.

### Exercise 2

If  $\{\underline{v}_1, \underline{v}_2, \dots, \underline{v}_n\}$  is a linearly independent set of vectors, prove that any subset of this set is also linearly independent.

### Exercise 3

If  $\{\underline{v}_1, \underline{v}_2, \dots, \underline{v}_n\}$  is a linearly dependent subset of a vector space  $V$ , prove that  $\{\underline{v}_1, \underline{v}_2, \dots, \underline{v}_n, \underline{w}\}$  is also linearly dependent, where  $\underline{w}$  is any element in  $V$ .

### Exercise 4

- (i) By choosing a suitable basis and finding its image, or otherwise describe the effect of the following mappings of  $R^2$  to  $R^2$ .

(a)  $T_1 : (x, y) \longmapsto (y, x)$

(b)  $T_2 : (x, y) \longmapsto \left( \frac{x}{\sqrt{2}} - \frac{y}{\sqrt{2}}, \frac{x}{\sqrt{2}} + \frac{y}{\sqrt{2}} \right)$

- (ii) Find an element in  $R^2$  which is unchanged under the mapping  $T_1$ .



*Exercise 5*

Which of the following mappings are morphisms? (Take the operations in the vector spaces to be the usual ones.)

- (i) The mapping of the set of polynomial functions of degree  $n$  or less to itself such that

$$T:p \longmapsto \text{the derived function of } p.$$

- (ii) The mapping of the set of all real functions, which are twice-differentiable at all points in  $R$ , to the set of all real functions, such that

$$T:f \longmapsto 2D^2f + Df + 3f$$

*Exercise 6*

Let  $L$  be a morphism from a vector space  $V$  to a vector space  $U$ . Show that the kernel of  $L$  is a vector subspace of  $V$ .

*Exercise 7*

We can map the vector space  $P_n$ , of all polynomial functions of degree  $n$  or less, to itself by the differentiation operator:

$$D:p \longmapsto p' \quad (p \in P_n).$$

We have already seen that this mapping is a morphism. What is the kernel? What significance does this have in integration?

## 5.8 Answers to Exercises

### Section 5.2

*Exercise 1*

The problem is caused when we add, say, the polynomial function  $x \longmapsto -x^n$  to the polynomial function  $x \longmapsto x^n + x^{n-1}$ . Both are of degree  $n$ , but their sum is the polynomial  $x \longmapsto x^{n-1}$ , which is of degree  $n-1$ , so  $+$  is not closed, i.e. axiom 1 is violated. A simple modification is to consider the set of polynomials of degree *less than or equal to*  $n$ . With the suggested operations, this set is indeed a vector space.

*Exercise 2*

- (i) No. For example, multiplication by a negative scalar takes us out of the set, i.e. axiom 5 is violated.



- (ii) Yes. All the axioms are satisfied. (In fact, all the points corresponding to the vectors lie on the line defined by the equation  $y + x = 0$ .)
- (iii) No. For example, if  $x_1 < x_2$  and  $\alpha < 0$ , then  $\alpha x_1 > \alpha x_2$ , i.e.  $\alpha \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$  does not belong to the given set, so axiom 5 is violated.
- (iv) Yes. All the axioms of a vector space are satisfied. (Each function has a graph which passes through the fixed point  $(k, 0)$ .)

### Exercise 3

- (i) The set of triples is linearly dependent; for instance,

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}.$$

- (ii) The set of functions is linearly independent, because  $\alpha f + \beta g = \underline{0}$  implies that  $\alpha x + \beta x^2 = 0$  for all values of  $x$ , and this is possible only if  $\alpha = \beta = 0$ .
- (iii) The set of triples is linearly independent.

$$\alpha_1 \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix} + \alpha_2 \begin{pmatrix} 0 \\ 3 \\ 0 \end{pmatrix} + \alpha_3 \begin{pmatrix} 0 \\ 0 \\ 5 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

that is

$$\begin{pmatrix} 2\alpha_1 \\ 3\alpha_2 \\ 5\alpha_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\text{or } \alpha_1 = \alpha_2 = \alpha_3 = 0$$

### Exercise 4

Using the axioms of a vector space, we can show that the given equation is equivalent to

$$(\alpha_1 - \beta_1)\underline{v}_1 + (\alpha_2 - \beta_2)\underline{v}_2 + \cdots + (\alpha_n - \beta_n)\underline{v}_n = \underline{0}.$$

Since the set of vectors  $\{\underline{v}_1, \underline{v}_2, \dots, \underline{v}_n\}$  is linearly independent, the coefficients of the vectors in the above equation are all zero, so

$$\alpha_1 - \beta_1 = \alpha_2 - \beta_2 = \cdots = \alpha_n - \beta_n = 0,$$

which proves the required result.



## Section 5.3

## Exercise 1

Any triple

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = (x_1 - x_2) \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + x_2 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + (x_3 - x_2) \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

Also the set of triples is linearly independent:

$$\alpha_1 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + \alpha_2 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + \alpha_3 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\text{i.e.} \quad \begin{pmatrix} \alpha_1 + \alpha_2 \\ \alpha_2 \\ \alpha_2 + \alpha_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

whence  $\alpha_1 = \alpha_2 = \alpha_3 = 0$ .

It follows that the given set of triples is a basis.

## Section 5.4

## Exercise 1

Straight lines which pass through the origin.

## Exercise 2

$$\underline{g}'_0: x \longmapsto 1 \quad (x \in R)$$

$$\underline{g}'_1: x \longmapsto -1 \quad (x \in R).$$

Although none of the base vectors is mapped to the zero vector,  $\{\underline{g}'_0, \underline{g}'_1, \underline{f}'_3\}$  is linearly *dependent*, since

$$1\underline{g}'_0 + (-1)\underline{g}'_1 + 0\underline{f}'_3 = \underline{f}'_0;$$

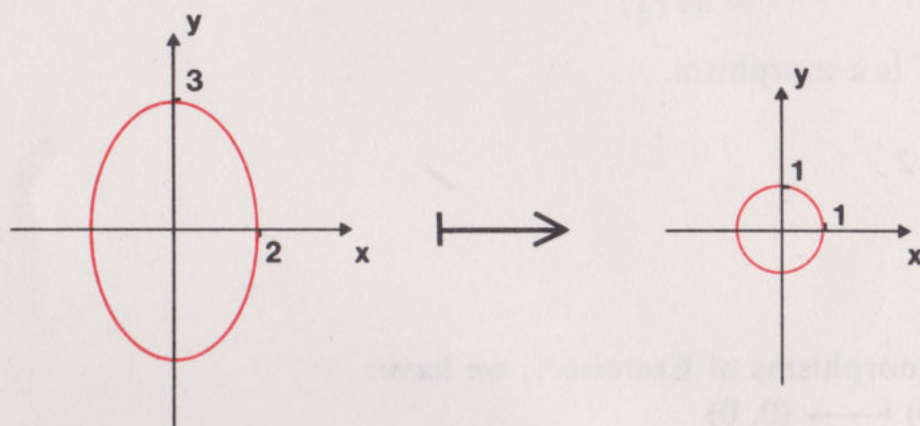
$\underline{g}'_0$  and  $\underline{g}'_1$  cannot both belong to the *same* basis because one is a scalar multiple of the other.



## Exercise 3

- (i)  $9(2u)^2 + 4(3v)^2 = 36$
- (ii)  $u^2 + v^2 = 1$
- (iii)  $\{(u, v): u^2 + v^2 = 1\}$
- (iv) the origin; 1

Thus the effect of the mapping is to transform the ellipse to the circle as shown below.



## Section 5.5

## Exercise 1

$$\begin{aligned}
 \text{(i)} \quad T((x_1, x_2) + (y_1, y_2)) &= T(x_1 + y_1, x_2 + y_2) \\
 &= (x_2 + y_2, x_1 + y_1) \\
 &= (x_2, x_1) + (y_2, y_1) \\
 &= T(x_1, x_2) + T(y_1, y_2),
 \end{aligned}$$

and

$$\begin{aligned}
 T(\alpha(x_1, x_2)) &= T(\alpha x_1, \alpha x_2) \\
 &= (\alpha x_2, \alpha x_1) \\
 &= \alpha(x_2, x_1) \\
 &= \alpha T(x_1, x_2),
 \end{aligned}$$

so  $T$  is a morphism.

$$\text{(ii)} \quad \alpha T(x_1, x_2) = \alpha(x_1^2, x_2^2) = (\alpha x_1^2, \alpha x_2^2),$$

and

$$T(\alpha(x_1, x_2)) = T(\alpha x_1, \alpha x_2) = (\alpha^2 x_1^2, \alpha^2 x_2^2)$$

Equation (2) is not satisfied, so  $T$  is not a morphism.



$$\begin{aligned}
 \text{(iii)} \quad T(\underline{x} + \underline{y}) &= q \cdot (\underline{x} + \underline{y}) \\
 &= q \cdot \underline{x} + q \cdot \underline{y} \quad (\cdot \text{ is distributive over } +) \\
 &= T(\underline{x}) + T(\underline{y})
 \end{aligned}$$

and

$$\begin{aligned}
 T(\alpha \underline{x}) &= q \cdot \alpha \underline{x} \\
 &= \alpha(q \cdot \underline{x}) \\
 &= \alpha T(\underline{x})
 \end{aligned}$$

so  $T$  is a morphism.

### Exercise 2

- (a)  $\underline{v}$
- (b)  $\underline{v}, \underline{v}_0$
- (c) zero

For the morphisms of Exercise 1, we have:

- (i)  $(0, 0) \longmapsto (0, 0)$
- (iii)  $\underline{0} \longmapsto 0$

### Exercise 3

- (i) As in the proof of the theorem, we need to check axioms 1, 4 and 5 only.

*Axiom 1* The elements of  $L(R^2)$  are of the form

$$(-a, a), a \in R,$$

and

$$(-a, a) + (-b, b) = (-(a + b), a + b),$$

and so axiom 1 is satisfied.

*Axiom 4* Consider the element  $(0, 0)$ .

$$L((0, 0)) = (0, 0),$$

so that  $(0, 0) \in L(R^2)$ , and axiom 4 is satisfied.

*Axiom 5*  $(-a, a)$  is any element of  $L(R^2)$

$$\alpha(-a, a) = (-\alpha a, \alpha a),$$

and so  $\alpha(-a, a) \in L(R^2)$ : axiom 5 is satisfied.

- (ii) The only axioms which *may* not be satisfied are 1, 4 and 5. Of these only 5 is not satisfied.

$$\alpha(x_1^2, x_2) = (\alpha x_1^2, \alpha x_2)$$



and if  $\alpha$  is negative,  $\alpha x_1^2$  is also negative, and so cannot be written as the square of a real number.

## Section 5.6

### Exercise 1

$$L(0, 1) = (-1, 1) \neq (0, 0)$$

and

$$L(2, 2) = (-2, 2) \neq (0, 0),$$

but

$$-2L(0, 1) + L(2, 2) = (0, 0)$$

Although neither vector maps to zero, the pair of linearly *independent* vectors maps to a pair of *dependent* vectors. So although the original vectors form a basis for  $R^2$ , their images do not.

### Exercise 2

- (i)  $\{(0, 0, x_3): x_3 \in R\}$ . Any element in this set is a scalar multiple of  $(0, 0, 1)$ .
- (ii)  $\{(x_1, x_2): x_1 + x_2 = 0 \text{ and } x_1 - 2x_2 = 0\}$

The pair of simultaneous equations

$$x_1 + x_2 = 0$$

$$x_1 - 2x_2 = 0$$

has the single solution  $x_1 = 0, x_2 = 0$ . Thus the kernel is the set  $\{(0, 0)\}$ .

The dimensions of the kernels are as follows:

- (i) 1. The dimension of the domain is 3, the dimension of its image set is 2, and  $3 - 2 = 1$ .
- (ii) 0. The dimension of both the domain and its image set is 2. Note that, by defining the dimension of  $\{0\}$  to be zero, we have ensured that (2) is satisfied when  $K = \{0\}$ .

### Exercise 3

Putting  $x = 0$  in both equations, gives

$$3y - z = 2$$

$$y + z = 1,$$

which have the solution

$$y = \frac{3}{4}, z = \frac{1}{4}$$

Thus one solution is

$$x = 0, y = \frac{3}{4}, z = \frac{1}{4}.$$



To find the kernel, we have to solve the equations

$$2x + 3y - z = 0$$

$$x + y + z = 0$$

If we put  $x = k$ , we get

$$3y - z = -2k$$

$$y + z = -k,$$

which have the solution

$$y = -\frac{3}{4}k, z = -\frac{1}{4}k;$$

so the kernel is the set

$$\{(k, -\frac{3}{4}k, -\frac{1}{4}k) : k \in R\}.$$

The complete solution is therefore

$$\{(k, \frac{3}{4} - \frac{3}{4}k, \frac{1}{4} - \frac{1}{4}k) : k \in R\}.$$

## Section 5.7

### Exercise 1

We check that the axioms of a vector space are satisfied.

1  $\underline{0} + \underline{0} = \underline{0}$ , since  $\underline{0}$  is the zero element of  $V$ , so  $\{\underline{0}\}$  is closed for addition.

4  $\underline{0} \in \{\underline{0}\}$ .

5  $\alpha \underline{0} = \underline{0} \in \{\underline{0}\}$  (proved in text).

The other axioms are automatically satisfied, since they are satisfied for all elements of  $V$ , and  $\underline{0} \in V$ ,

Hence  $\{\underline{0}\}$  is a (real) vector space.

### Exercise 2

Suppose, in contradiction to what we want to prove, that the subset  $\{\underline{v}_1, \underline{v}_2, \dots, \underline{v}_k\}$  is linearly dependent; then there are numbers  $\alpha_1, \alpha_2, \dots, \alpha_k$  (not all zero) such that

$$\alpha_1 \underline{v}_1 + \alpha_2 \underline{v}_2 + \dots + \alpha_k \underline{v}_k = \underline{0}.$$

Therefore

$$\alpha_1 \underline{v}_1 + \alpha_2 \underline{v}_2 + \dots + \alpha_k \underline{v}_k + 0 \underline{v}_{k+1} + \dots + 0 \underline{v}_n = \underline{0}.$$

But not *all* the  $\alpha_1, \dots, \alpha_n$  are zero, and hence the set of vectors  $\{\underline{v}_1, \underline{v}_2, \dots, \underline{v}_n\}$  is linearly dependent—which is a contradiction.



## Exercise 3

If  $\{\underline{v}_1, \underline{v}_2, \dots, \underline{v}_n\}$  is linearly dependent, then there are numbers  $\alpha_1, \alpha_2, \dots, \alpha_n$ , not all zero, such that

$$\alpha_1 \underline{v}_1 + \alpha_2 \underline{v}_2 + \dots + \alpha_n \underline{v}_n = \underline{0}.$$

Hence

$$\alpha_1 \underline{v}_1 + \alpha_2 \underline{v}_2 + \dots + \alpha_n \underline{v}_n + 0 \underline{w} = \underline{0}.$$

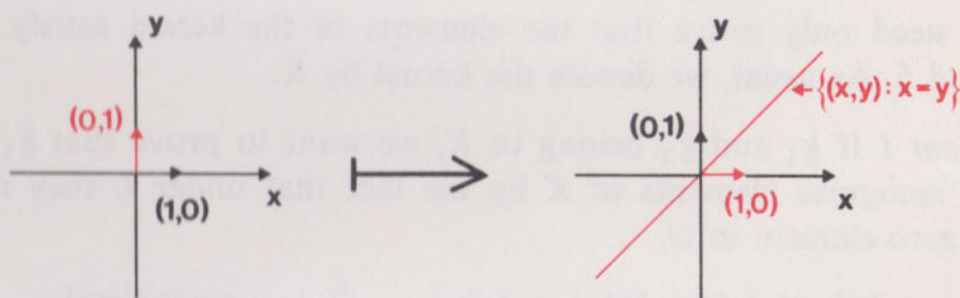
Not all the coefficients in this last equation are zero, and so we have proved the required result.

## Exercise 4

(i) If we choose the basis  $\{(1, 0), (0, 1)\}$ , we have:

$$T_1: (1, 0) \mapsto (0, 1)$$

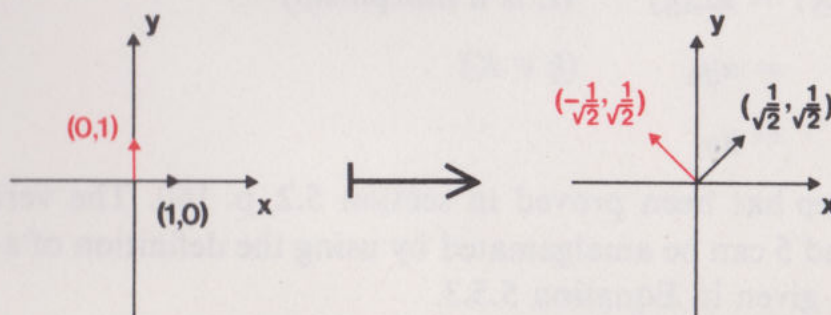
$$T_1: (0, 1) \mapsto (1, 0)$$



The effect of  $T_1$  is to reflect the points of the plane in the line with equation  $y = x$  (we are simply interchanging the  $x$  and  $y$  co-ordinates).

$$(b) T_2: (1, 0) \mapsto \left( \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right)$$

$$T_2: (0, 1) \mapsto \left( -\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right)$$





The effect of  $T_2$  is to rotate the plane about the origin through an angle  $\frac{\pi}{4}$ . (If you are not convinced, find the images of some more points.)

- (ii) Any scalar multiple of  $(1, 1)$  is an invariant element under  $T_1$ . In terms of geometric vectors, any vector in the direction of the line with equation  $x = y$  is invariant. All such geometric vectors are scalar multiples of  $\underline{i} + \underline{j}$ .

### Exercise 5

- (i)  $T$  is a morphism } These results follow directly from the properties  
(ii)  $T$  is a morphism } of  $D$ .

### Exercise 6

We need only prove that the elements of the kernel satisfy axioms 1, 4 and 5. As usual, we denote the kernel by  $K$ .

*Axiom 1* If  $\underline{k}_1$  and  $\underline{k}_2$  belong to  $K$ , we want to prove that  $\underline{k}_1 + \underline{k}_2 \in K$ . We recognize elements of  $K$  by the fact that under  $L$  they map to  $\underline{u}_0$ , the zero element in  $U$ .

$$\begin{aligned} L(\underline{k}_1 + \underline{k}_2) &= L(\underline{k}_1) + L(\underline{k}_2) && (L \text{ is a morphism}) \\ &= \underline{u}_0 + \underline{u}_0 && (\underline{k}_1, \underline{k}_2 \in K) \\ &= \underline{u}_0 && (\text{axiom 4 for } U) \end{aligned}$$

Therefore,  $\underline{k}_1 + \underline{k}_2 \in K$ .

*Axiom 4* We have already shown (Exercise 5.5.2) that  $L(\underline{v}_0) = \underline{u}_0$ . Therefore  $\underline{v}_0 \in K$ .

*Axiom 5* Let  $\underline{k} \in K$ , then

$$\begin{aligned} L(\alpha \underline{k}) &= \alpha L(\underline{k}) && (L \text{ is a morphism}) \\ &= \alpha \underline{u}_0 && (\underline{k} \in K) \\ &= \underline{u}_0 \end{aligned}$$

This last step has been proved in section 5.2, p. 160. The verification of axioms 1 and 5 can be amalgamated by using the definition of a morphism in the form given in Equation 5.5.3.



*Exercise 7*

The kernel is the set of all polynomial functions which map to the zero function ( $x \mapsto 0$  ( $x \in R$ )); that is, the set of all constant functions,

$$\{f: x \mapsto k(x \in R), k \in R\}.$$

The problem of integration is that of solving equations of the form

$$D(p) = f,$$

where  $f$  is given.

Since the kernel contains an infinite number of elements, the integration problem has an infinite number of solutions. If  $p$  is one solution, then the set of all solutions is

$$\{p + f: f: x \mapsto k(x \in R), k \in R\}.$$



# CHAPTER 6 MATRICES

## 6.0 Introduction

In this chapter we introduce a new idea—that of a *matrix*. A *matrix* is simply a rectangular (or square) array of numbers such as you might see on a bus timetable or on the board outside a cinema, showing the starting times of the films. Such arrays of numbers arise when we try to quantify almost any investigation.

For example, the following two arrays might represent a survey of the voting inclinations in two constituencies.

	Party X	Party Y	Party Z	Don't Know
Men	30%	35%	10%	25%
Women	25%	50%	15%	10%

	Party X	Party Y	Party Z	Don't Know
Men	50%	40%	5%	5%
Women	40%	40%	2%	18%

There are various ways of combining arrays. In this example, adding corresponding entires and dividing by 2 to give

	Party X	Party Y	Party Z	Don't Know
Men	40%	37½%	7½%	15%
Women	32½%	45%	8½%	14%

might be a meaningful thing to do, to give a sort of “average” voting inclination.

In the first two sections of this chapter we show how morphisms between vector spaces can be represented by matrices. This correspondence between mappings and matrices leads us to define some ways of combining matrices. We may want to define special combinations for special applications, but the ones we define here are the most widely used.



## 6.1 Linear Equations

In the previous chapter we discussed several examples of mappings from  $R^3$  to  $R^2$  or  $R^3$  to  $R^3$ , and we saw in Example 5.6.3 that these mappings are connected with simultaneous equations. It is fairly clear that three equations, each expressing one of  $x'_1$ ,  $x'_2$  and  $x'_3$  in terms of  $x_1$ ,  $x_2$  and  $x_3$ , will tell us how to map  $(x_1, x_2, x_3)$  to  $(x'_1, x'_2, x'_3)$ . If these equations hold for all real values of  $x_1, x_2, x_3$ , then they will define a mapping from  $R^3$  to  $R^3$ . For example, the three equations

$$x'_1 = \sin x_1$$

$$x'_2 = \sin x_2$$

$$x'_3 = \sin x_3$$

Equations (1)

define such a mapping, as do the equations

$$x'_1 = x_1 + x_2 + x_3$$

$$x'_2 = 2x_1 - x_2$$

$$x'_3 = 2x_2 + 3x_3$$

Equations (2)

where the right-hand sides are linear combinations of  $x_1, x_2$  and  $x_3$ .

### Exercise 1

Do Equations (1) define a morphism from  $R^3$  to  $R^3$ ?

Do Equations (2) define a morphism from  $R^3$  to  $R^3$ ?

It seems that *any* set of three equations of the form:

$$x'_1 = a_1x_1 + a_2x_2 + a_3x_3$$

$$x'_2 = b_1x_1 + b_2x_2 + b_3x_3$$

$$x'_3 = c_1x_1 + c_2x_2 + c_3x_3$$

Equations (3)

where the  $a$ 's,  $b$ 's and  $c$ 's are real numbers, defines a morphism from  $R^3$  to  $R^3$ . A set of equations of this form is called a set of **linear** equations, and our suggestion can be put in more general terms as in the following proposition.

Any set of  $m$  linear equations, each expressing one of  $m$  variables

$$x'_1, x'_2, \dots, x'_m \text{ in terms of } n \text{ variables } x_1, x_2, \dots, x_n$$

defines a morphism of  $R^n$  to  $R^m$ .



The proposition is easily confirmed, for if the equations are

$$x'_1 = a_1x_1 + a_2x_2 + \cdots + a_nx_n$$

$$x'_2 = b_1x_1 + b_2x_2 + \cdots + b_nx_n$$

$$\dots\dots\dots$$

$$\dots\dots\dots$$

$$x'_m = q_1x_1 + q_2x_2 + \cdots + q_nx_n$$

Equations (4)

they certainly define a *mapping*  $T$  of  $R^n$  to  $R^m$ :

$$T:(x_1, x_2, \dots, x_n) \longmapsto (x'_1, x'_2, \dots, x'_m).$$

For the variables  $x_1, x_2, \dots, x_n$  can each take any real value, and so  $(x_1, x_2, \dots, x_n)$  can be any element of  $R_n$ . We have to prove that  $T$  is a morphism, i.e. that (see section 5.5)

$$T(\alpha x_1, \alpha x_2, \dots, \alpha x_n) = \alpha T(x_1, x_2, \dots, x_n)$$

where  $\alpha$  is any real number, and

$$\begin{aligned} T(x_1 + y_1, x_2 + y_2, \dots, x_n + y_n) \\ = T(x_1, x_2, \dots, x_n) + T(y_1, y_2, \dots, y_n) \end{aligned}$$

The first statement follows from the fact that, for example,

$$\begin{aligned} a_1(\alpha x_1) + a_2(\alpha x_2) + \cdots + a_n(\alpha x_n) \\ = \alpha(a_1x_1 + a_2x_2 + \cdots + a_nx_n) \end{aligned}$$

a process which can be applied to each of the equations.

Similarly, we have, for example,

$$\begin{aligned} a_1(x_1 + y_1) + a_2(x_2 + y_2) + \cdots + a_n(x_n + y_n) \\ = (a_1x_1 + a_2x_2 + \cdots + a_nx_n) + (a_1y_1 + a_2y_2 + \cdots + a_ny_n). \end{aligned}$$

The second statement follows from an application of this process to each component of  $T(x_1 + y_1, \dots, x_n + y_n)$ .

We thus have the result that every set of linear equations of the form of Equations (4) defines a *morphism*.

This in itself would not be so interesting, if it were not for the fact that we can represent *any* morphism from  $R^n$  to  $R^m$  by such a set of equations. We can do this as follows.



We take as a basis for  $R^n$  the set of  $n$  vectors:

$$\underline{e}_1 = (1, 0, 0, 0, \dots, 0),$$

$$\underline{e}_2 = (0, 1, 0, 0, \dots, 0),$$

$$\underline{e}_3 = (0, 0, 1, 0, \dots, 0),$$

$\dots$

$$\underline{e}_n = (0, 0, 0, 0, \dots, 1).$$

An arbitrary element of  $R^n$  can be written:

$$\underline{x} = (x_1, x_2, x_3, \dots, x_n) = x_1 \underline{e}_1 + x_2 \underline{e}_2 + x_3 \underline{e}_3 + \dots + x_n \underline{e}_n.$$

If  $T$  is a morphism from  $R^n$  to  $R^m$ , then

$$T(\underline{x}) = T(x_1 \underline{e}_1 + x_2 \underline{e}_2 + \dots + x_n \underline{e}_n) = x_1 T(\underline{e}_1) + \dots + x_n T(\underline{e}_n).$$

We see that the base vectors tell us the whole story. Each of the base vectors has an image in  $R^m$ ; suppose

$$T(\underline{e}_1) = (a_1, b_1, \dots, q_1)$$

$$T(\underline{e}_2) = (a_2, b_2, \dots, q_2)$$

$\dots$

$$T(\underline{e}_n) = (a_n, b_n, \dots, q_n)$$

Then

$$\begin{aligned} T(\underline{x}) &= x_1(a_1, b_1, \dots, q_1) \\ &+ x_2(a_2, b_2, \dots, q_2) \\ &\dots \\ &+ x_n(a_n, b_n, \dots, q_n) \\ &= (a_1 x_1 + a_2 x_2 + \dots + a_n x_n) \underline{e}_1 \\ &+ (b_1 x_1 + b_2 x_2 + \dots + b_n x_n) \underline{e}_2 \\ &+ \dots \\ &+ (q_1 x_1 + q_2 x_2 + \dots + q_n x_n) \underline{e}_m \end{aligned}$$

In this calculation we have first expressed  $\underline{x}$  as a linear combination of the basis  $\underline{e}_1, \underline{e}_2, \dots, \underline{e}_n$  and then used the fact that  $T$  is a morphism to calculate the image of  $\underline{x}$  as a linear combination of the images of the  $\underline{e}$ 's. But, in theory, we can also calculate the image of  $\underline{x}$  as an element of  $R^m$  directly. Suppose, therefore, that

$$T(\underline{x}) = T(x_1, x_2, \dots, x_n) = (x'_1, x'_2, \dots, x'_m)$$



Then we can write

$$T(\underline{x}) = x'_1 \underline{e}_1 + x'_2 \underline{e}_2 + \cdots + x'_m \underline{e}_m$$

We now have two expressions for  $T(\underline{x})$  as a linear combination of the  $\underline{e}$ 's. But a vector cannot be expressed in two different ways as a linear combination of a set of linearly independent vectors (see Exercise 5.2.4), so that we have

$$x'_1 = a_1 x_1 + a_2 x_2 + \cdots + a_n x_n$$

$$x'_2 = b_1 x_1 + b_2 x_2 + \cdots + b_n x_n$$

Equations 5

$$x'_m = q_1 x_1 + q_2 x_2 + \cdots + q_n x_n$$

Thus, the morphism  $T$  can be represented by Equations (5).

There is a one-one correspondence between sets of linear equations and morphisms from  $R^n$  to  $R^m$ . Any set of linear equations of the form (5) defines a morphism from  $R^n$  to  $R^m$ , and any morphism from  $R^n$  to  $R^m$  defines a set of linear equations of the form (5).

We mentioned earlier that a morphism between vector spaces is often called a *linear transformation*. We now see why—the *transformation* refers to the mapping part, and *linear* refers to the properties we have just discussed. It is from here also that we get the term *Linear Algebra*.

In the next chapter we shall take up the topic of sets of linear equations in much more detail. In the meantime we shall use the ideas we have just been discussing to introduce the topic of *matrices*, which is of great importance in *Linear Algebra* and mathematics generally, and which we shall use when we discuss linear equations in detail.

## 6.2 Matrices

We have seen that Equations 6.1.5 define a mapping from  $R^n$  to  $R^m$  (given bases for  $R^n$  and  $R^m$ ). Now  $(x_1, x_2, \dots, x_n)$  and  $(x'_1, x'_2, \dots, x'_m)$  are simply arbitrary elements in the respective vector spaces; the equations themselves are adequately specified by the array of numbers

$$a_1 \quad a_2 \quad \cdots \quad a_n$$

$$b_1 \quad b_2 \quad \cdots \quad b_n$$

$$\cdots$$

$$\cdots$$

$$q_1 \quad q_2 \quad \cdots \quad q_n$$

Such an array of numbers is called a **matrix**.



Often we want to talk about this array as an entity in itself, as opposed to a collection of elements. To do this we either put parentheses round the array or refer to it by a single letter. Sometimes we do both, thus we write

$$A \text{ or } \begin{pmatrix} a_1 & a_2 & \cdots & a_n \\ b_1 & b_2 & \cdots & b_n \\ \cdots & & & \\ q_1 & q_2 & \cdots & q_n \end{pmatrix} \begin{array}{l} \leftarrow \\ \leftarrow \\ \leftarrow \\ \leftarrow \end{array} \left. \vphantom{\begin{pmatrix} a_1 \\ b_1 \\ \cdots \\ q_1 \end{pmatrix}} \right\} m \text{ rows}$$

$$\begin{array}{c} \uparrow \quad \uparrow \quad \cdots \quad \uparrow \\ \underbrace{\hspace{1.5cm}} \\ n \text{ columns} \end{array}$$

As we have seen, a matrix with  $m$  rows and  $n$  columns defines a mapping from  $R^n$  to  $R^m$  (the bases being given or understood), and the mapping is generally referred to by the name of the matrix.

### Example 1

- (i) The mapping of  $R^2$  to  $R^2$  defined by

$$x' = 2x + 3y$$

$$y' = 3x - y$$

is represented by the matrix

$$\begin{pmatrix} 2 & 3 \\ 3 & -1 \end{pmatrix}$$

- (ii) The matrix

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 0 & 6 \end{pmatrix}$$

represents the mapping of  $R^3$  to  $R^2$  defined by

$$x'_1 = x_1 + 2x_2 + 3x_3$$

$$x'_2 = 4x_1 + 0x_2 + 6x_3$$

Two matrices are said to be **equal** if they represent the same mapping (with respect to bases which are given or understood). Thus matrices  $A$  and  $B$  are equal if they have the same number of rows and columns (they must represent mappings with the same domain and codomain), and the corresponding elements are equal. If  $A$  and  $B$  are equal, we write  $A = B$ .



For example,

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \neq \begin{pmatrix} 2 & 4 \\ 6 & 8 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 0 & 0 & 0 \end{pmatrix} \neq \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{pmatrix}$$

$$(1 \ 2 \ 3 \ 4) \neq \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Since a matrix is an alternative way of specifying a morphism, we should be able to tell something about the morphism just by looking at the matrix. Investigating the properties of matrices is equivalent to investigating morphisms between vector spaces.

### 6.3 Combining Matrices

We can *represent* a set of equations by a matrix, but we can also go one better and *rewrite* the equations in terms of this matrix. Consider, for example, the equations

$$x'_1 = 3x_1 + 2x_2 + 1x_3$$

Equation (1)

$$x'_2 = 1x_1 + 1x_2 + 3x_3$$

They specify a mapping from  $R^3$  to  $R^2$  under which

$$(x_1, x_2, x_3) \longmapsto (x'_1, x'_2),$$

and are represented by the matrix

$$A = \begin{pmatrix} 3 & 2 & 1 \\ 1 & 1 & 3 \end{pmatrix}$$

We can write the equations in the following way:

$$(x'_1, x'_2) = A(x_1, x_2, x_3),$$

but there are plenty of other ways in which we could write them. We shall



choose a way which leads to results consistent with the existing literature on matrices. We write

$$(x'_1, x'_2) \text{ as } \begin{pmatrix} x'_1 \\ x'_2 \end{pmatrix}$$

and

$$(x_1, x_2, x_3) \text{ as } \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix};$$

that is, we write the elements of  $R^2$  and  $R^3$  as arrays (or lists) of numbers, rather than in co-ordinate form. These lists are just special kinds of matrices, matrices with *one* column. So we can rewrite the equations in matrix notation as

$$\begin{pmatrix} x'_1 \\ x'_2 \end{pmatrix} = A \square \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \quad \text{Equation (2)}$$

where  $\square$  is a combination of matrices which we have to define to make Equation (2) equivalent to Equations (1).

From our definition of equality of matrices,  $A \square \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$  must be a matrix

with *one* column and *two* rows, and the first element must be  $x'_1$  and the second  $x'_2$ . But  $x'_1$  and  $x'_2$  are given by Equations (1), so

$$A \square \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 3x_1 + 2x_2 + x_3 \\ x_1 + x_2 + 3x_3 \end{pmatrix}$$

Thus

$$\begin{pmatrix} 3 & 2 & 1 \\ 1 & 1 & 3 \end{pmatrix} \square \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 3x_1 + 2x_2 + 1x_3 \\ 1x_1 + 1x_2 + 3x_3 \end{pmatrix}$$



and for a general mapping from  $R^3$  to  $R^2$ :

$$\begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix} \square \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} = \begin{pmatrix} a\alpha + b\beta + c\gamma \\ d\alpha + e\beta + f\gamma \end{pmatrix}$$

The *first element* in the third matrix comes from the *first row* of the first matrix, and is obtained by multiplying each element of that row with the corresponding element in the second matrix and adding together the products thus formed. The *second element* comes from the *second row* of the first matrix by applying the same procedure.

For example,

$$\begin{pmatrix} 1 & 2 & 3 \\ -2 & 3 & -4 \end{pmatrix} \square \begin{pmatrix} 1 \\ -2 \\ 3 \end{pmatrix} = \begin{pmatrix} 1 \times 1 + 2 \times (-2) + 3 \times 3 \\ (-2) \times 1 + 3 \times (-2) + (-4) \times 3 \end{pmatrix} \\ = \begin{pmatrix} 6 \\ -20 \end{pmatrix}$$

We can extend this operation to deal with longer lists and matrices with more rows and columns. This extension corresponds to manipulating equations specifying mappings from  $R^n$  to  $R^m$ , where the column corresponding to an element of  $R^n$  has  $n$  elements

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \text{ and the left-hand matrix is } \begin{pmatrix} a_1 & a_2 & \cdots & a_n \\ b_1 & b_2 & \cdots & b_n \\ \vdots & \vdots & \ddots & \vdots \\ q_1 & q_2 & \cdots & q_n \end{pmatrix} \left. \vphantom{\begin{pmatrix} a_1 & a_2 & \cdots & a_n \\ b_1 & b_2 & \cdots & b_n \\ \vdots & \vdots & \ddots & \vdots \\ q_1 & q_2 & \cdots & q_n \end{pmatrix}} \right\} m \text{ rows}$$

This matrix must, of course, have the same number of columns as there are elements in the list matrix corresponding to an element of  $R^n$ , because of the form of the linear equations.

### Example 1

$$(i) \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} \square \begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix} = \begin{pmatrix} 1 \times 1 + 2 \times (-1) + 3 \times 2 \\ 4 \times 1 + 5 \times (-1) + 6 \times 2 \\ 7 \times 1 + 8 \times (-1) + 9 \times 2 \end{pmatrix} = \begin{pmatrix} 5 \\ 11 \\ 17 \end{pmatrix}$$



$$(ii) \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix} \square \begin{pmatrix} 7 \\ 8 \end{pmatrix} = \begin{pmatrix} 23 \\ 53 \\ 83 \end{pmatrix}$$

(iii) The set of equations

$$2x_1 + 3x_2 = 2$$

$$1x_1 - 2x_2 = 3$$

$$3x_1 + 1x_2 = 4$$

can be written in the form

$$\begin{pmatrix} 2 & 3 \\ 1 & -2 \\ 3 & 1 \end{pmatrix} \square \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix}$$

### Exercise 1

Calculate

$$(i) \begin{pmatrix} 1 & -1 \\ 2 & 0 \\ 3 & 4 \end{pmatrix} \square \begin{pmatrix} 3 \\ -2 \end{pmatrix} \quad (ii) \begin{pmatrix} 1 & 2 & 3 \\ -1 & 0 & 4 \end{pmatrix} \square \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$

So far in this section we have combined rather special matrices; the second matrix has always had just a single column. There are a number of ways of motivating methods of combining more general matrices, but in our context, where we have associated matrices with vector space morphisms, the most natural way is to derive our matrix combinations from the combination of morphisms. Remembering that morphisms are functions (with special properties), we can consider the relevant combinations of functions as described in section 1.3.

As a first step we consider the composition of functions and show that if  $T_1$  is a morphism of the vector space  $V$  to the vector space  $U$ , and  $T_2$  is a morphism of  $U$  to the vector space  $W$ , then  $T_2 \circ T_1$  is a morphism of  $V$  to  $W$ . With the usual notation, we have

$$\begin{aligned} T_2 \circ T_1(\alpha_1 \underline{v}_1 + \alpha_2 \underline{v}_2) &= T_2(T_1(\alpha_1 \underline{v}_1 + \alpha_2 \underline{v}_2)) \\ &= T_2(\alpha_1 T_1(\underline{v}_1) + \alpha_2 T_1(\underline{v}_2)) \\ &\quad \text{(since } T_1 \text{ is a morphism)} \end{aligned}$$



$$\begin{aligned}
&= \alpha_1 T_2(T_1(v_1)) + \alpha_2 T_2(T_1(v_2)) \\
&\quad \text{(since } T_2 \text{ is a morphism)} \\
&= \alpha_1 T_2 \circ T_1(v_1) + \alpha_2 T_2 \circ T_1(v_2)
\end{aligned}$$

i.e. Equation 5.5.3 holds, so  $T_2 \circ T_1$  is a morphism.

It follows that we can define a combination of matrices corresponding to the composition of morphisms.

Consider, for example, the morphisms  $T_1$  and  $T_2$  of  $R^2$  to  $R^2$  defined by

$$T_1:(x_1, x_2) \longmapsto (1x_1 + 1x_2, 2x_1 + 1x_2)$$

$$T_2:(x_1, x_2) \longmapsto (1x_1 + 2x_2, 3x_1 + 2x_2)$$

Under the composite mapping  $T_2 \circ T_1$  we have

$$T_2 \circ T_1:(x_1, x_2) \longmapsto T_2(T_1(x_1, x_2))$$

and

$$\begin{aligned}
T_2(T_1(x_1, x_2)) &= T_2(1x_1 + 1x_2, 2x_1 + 1x_2) \\
&= (1(x_1 + x_2) + 2(2x_1 + x_2), 3(x_1 + x_2) \\
&\quad + 2(2x_1 + x_2)) \\
&= (5x_1 + 3x_2, 7x_1 + 5x_2)
\end{aligned}$$

Writing

$$T_1(x_1, x_2) = (x'_1, x'_2) \quad \text{and} \quad T_2(x'_1, x'_2) = (x''_1, x''_2),$$

we have:

$$x'_1 = 1x_1 + 1x_2$$

$$x'_2 = 2x_1 + 1x_2$$

Equations (3)

and

$$x''_1 = 1x'_1 + 2x'_2$$

$$x''_2 = 3x'_1 + 2x'_2$$

Equations (4)

Substituting from Equations (3) into Equations (4), we get

$$x''_1 = 5x_1 + 3x_2$$

$$x''_2 = 7x_1 + 5x_2$$

Equations (5)

as before.



In terms of matrices, the matrix representation for  $T_1$  is

$$A_1 = \begin{pmatrix} 1 & 1 \\ 2 & 1 \end{pmatrix}$$

(see Equations (3)), and the matrix representation for  $T_2$  is

$$A_2 = \begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix}$$

(see Equations (4)). And we have shown that the matrix representation for  $T_2 \circ T_1$  is

$$A_3 = \begin{pmatrix} 5 & 3 \\ 7 & 5 \end{pmatrix}$$

(see Equations (5)).

If we use the symbol  $*$  to represent the operation of combining matrices which corresponds to the composition of mappings, we have

$$A_2 * A_1 = A_3$$

i.e.

$$\begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix} * \begin{pmatrix} 1 & 1 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} 5 & 3 \\ 7 & 5 \end{pmatrix} \quad \text{Equations (6)}$$

We now know how to combine  $A_1$  and  $A_2$  in this case, but how can we work out  $B * A$  for *any* matrices  $B$  and  $A$ ? Indeed, does  $B * A$  *make sense* for *any* matrices  $B$  and  $A$ ?

First of all, how do we calculate  $B * A$ ? Looking at Equation (6), we see that the element in the *first* row and *first* column of  $A_3$  is 5, and

$$5 = 1 \times 1 + 2 \times 2,$$

and this expression is obtained by multiplying the elements in the *first* row of  $A_2$  by the corresponding elements in the *first* column of  $A_1$  and adding together the products.

$$\begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix} * \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 5 \\ \end{pmatrix}$$

Combining rows of  $A_2$  with columns of  $A_1$  in this way, we get the other elements of  $A_1$ .

$$\begin{pmatrix} 3 \\ 2 \end{pmatrix} * \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 7 \\ \end{pmatrix}$$



$$\begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix} * \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ 5 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix} * \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ 5 \end{pmatrix}$$

In the general case of matrices with two rows and two columns, we define  $*$  by

$$\begin{pmatrix} e & f \\ g & h \end{pmatrix} * \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} ea + fc & eb + fd \\ ga + hc & gb + hd \end{pmatrix}$$

### Exercise 2

If

$$x'_1 = ax_1 + bx_2$$

$$x'_2 = cx_1 + dx_2 \tag{i}$$

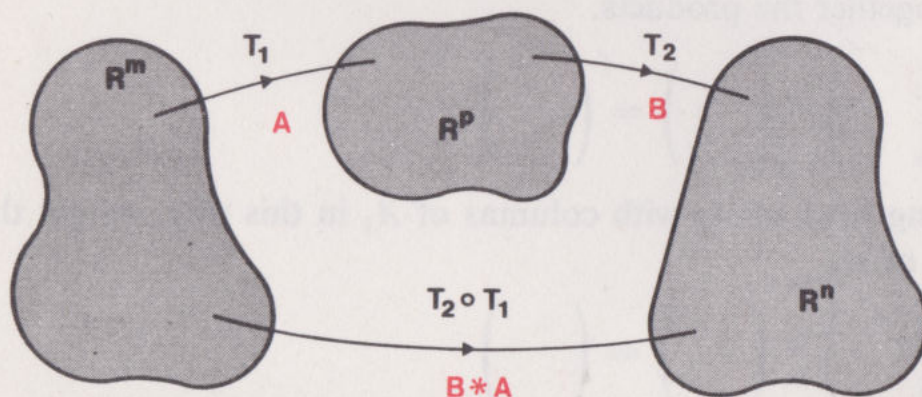
and

$$x''_1 = ex'_1 + fx'_2$$

$$x''_2 = gx'_1 + hx'_2 \tag{ii}$$

find  $x''_1$  and  $x''_2$  in terms of  $x_1$  and  $x_2$ , and compare your answer with the definition of  $*$ .

We can extend the definition of  $*$  to more general matrices. So far we have combined matrices with two rows and two columns—corresponding to combining a mapping from  $R^2$  to  $R^2$  with another mapping from  $R^2$  to  $R^2$ . It makes sense to combine mappings other than these; a morphism  $T_1$  from  $R^m$  to  $R^p$ , say, can be followed by a morphism  $T_2$  from  $R^p$  to  $R^n$ , say.





We know that  $T_1$  will be represented by a matrix with  $m$  columns and  $p$  rows, and  $T_2$  will be represented by a matrix with  $p$  columns and  $n$  rows. So we extend the definition of  $*$  to form  $B * A$  for all sorts of matrices provided the number of columns in  $B$  is the same as the number of rows in  $A$ . Since  $T_2 \circ T_1$  is a morphism from  $R^m$  to  $R^n$ , the combined matrix will have  $n$  rows and  $m$  columns, and in general  $B * A$  will be a matrix with the same number of rows as  $B$  and the same number of columns as  $A$ .

$$\begin{array}{c} \text{rows } n \\ \left( \begin{array}{c} p \\ \text{columns} \end{array} \right) \end{array} * \begin{array}{c} \text{rows } p \\ \left( \begin{array}{c} m \\ \text{columns} \end{array} \right) = \begin{array}{c} \text{rows } n \\ \left( \begin{array}{c} m \\ \text{columns} \end{array} \right)$$

We shall not follow through all the details, but illustrate how we might proceed in a simple example.

### Example 2

$$\text{If } B = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix} \text{ and } A = \begin{pmatrix} 1 \\ 2 \end{pmatrix},$$

how do we calculate  $B * A$ ?  $A$  represents the mapping from  $R^1$  to  $R^2$  specified by the equations

$$x'_1 = 1x_1$$

$$x'_2 = 2x_1$$

Equations (7)

and  $B$  represents the mapping from  $R^2$  to  $R^3$  specified by the equations

$$x''_1 = 1x'_1 + 2x'_2$$

$$x''_2 = 3x'_1 + 4x'_2$$

$$x''_3 = 5x'_1 + 6x'_2$$

Equations (8)

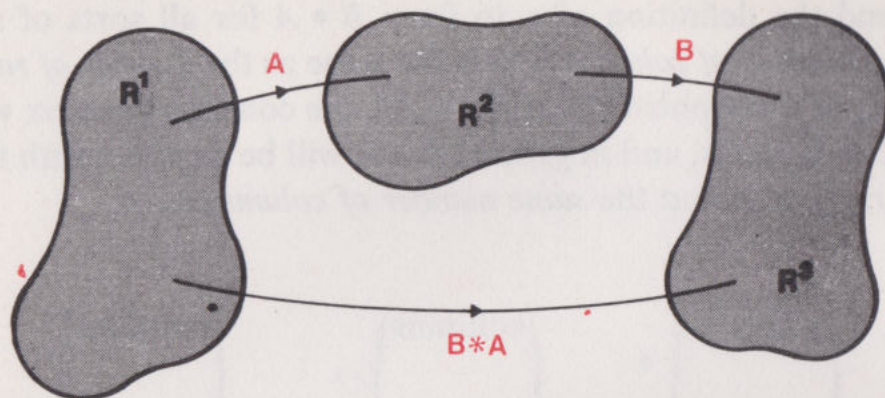
Substituting Equations (7) into Equations (8), we find that the composite mapping has matrix

$$B * A = \begin{pmatrix} 5 \\ 11 \\ 17 \end{pmatrix}$$

This is a matrix with three rows and one column; we expect this because



the result of mapping  $R^1$  to  $R^2$  (matrix  $A$ ) and then  $R^2$  to  $R^3$  (matrix  $B$ ) is a mapping from  $R^1$  to  $R^3$  (matrix  $B * A$ ).



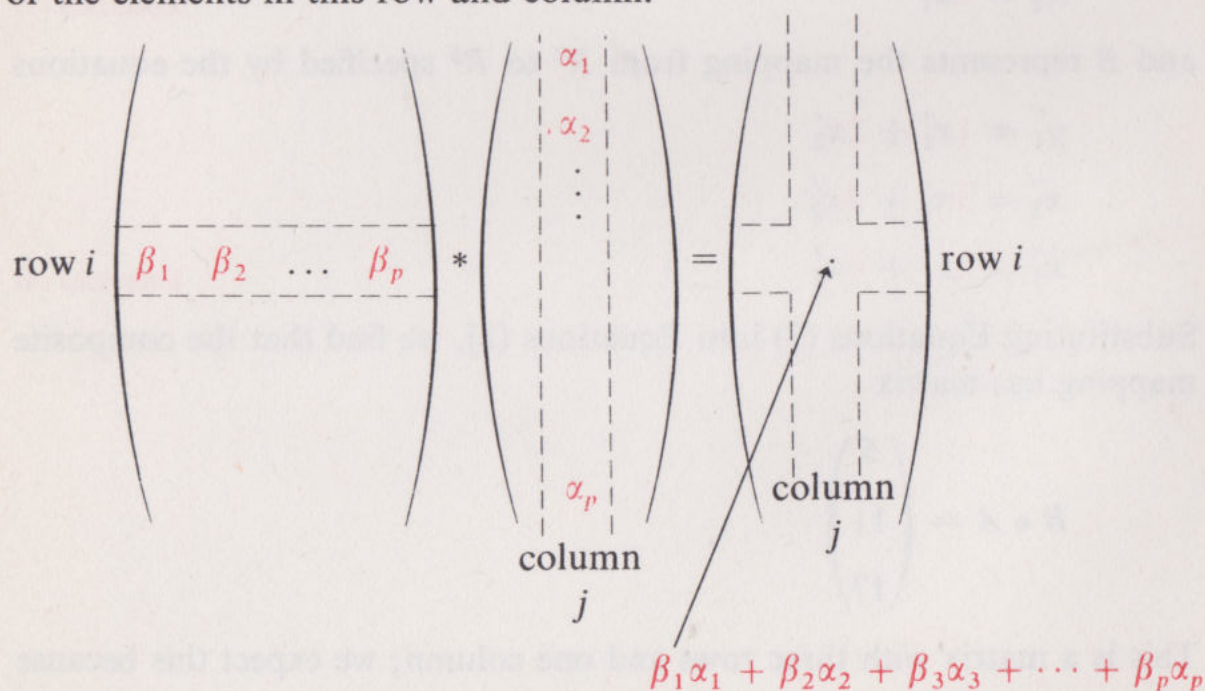
The elements in the matrix  $B * A$  are calculated by exactly the same method as before: the element 5 comes from the *first* row of  $B$  and the *first* (and only) column of  $A$ :  $5 = 1 \times 1 + 2 \times 2$ . The element 11 comes from the *second* row of  $B$  and the *first* (and only) column of  $A$ :  $11 = 3 \times 1 + 4 \times 2$ . And  $17 = 5 \times 1 + 6 \times 2$ .

If you look back at page 207 where we discussed the operation  $\square$ , you will see that  $\square$  is just a special case of  $*$ . We can define  $*$  for general matrices (to correspond to the composition of morphisms) as follows.

The combination  $B * A$  is defined for any matrices  $B$  and  $A$ , provided that the *number of columns in  $B$*  is the same as the *number of rows in  $A$* . If

$$B * A = C,$$

then the element in row  $i$  and column  $j$  of  $C$  is obtained from row  $i$  of  $B$  and column  $j$  of  $A$ ; it is calculated by adding together the termwise products of the elements in this row and column.





If  $B$  has  $m$  rows and  $p$  columns and  $A$  has  $p$  rows and  $n$  columns, then  $C = B * A$  will have  $m$  rows and  $n$  columns.

### Example 3

$$(i) \begin{pmatrix} 1 & -2 \\ -1 & 3 \\ 2 & 4 \end{pmatrix} * \begin{pmatrix} 1 & 3 & -4 \\ -1 & -2 & 1 \end{pmatrix} = \begin{pmatrix} 3 & 7 & -6 \\ -4 & -9 & 7 \\ -2 & -2 & -4 \end{pmatrix}$$

$$(ii) (1 \ 2 \ 3 \ 4 \ 5) * \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = (15)$$

$$(iii) \begin{pmatrix} 1 & -2 \\ -1 & 3 \\ 2 & 4 \end{pmatrix} * \begin{pmatrix} 1 & 3 & -4 \\ -1 & -2 & 1 \\ 1 & 2 & 3 \end{pmatrix} \text{ is not defined, because}$$

the number of columns in the left-hand matrix is *not* the same as the number of rows in the right-hand matrix.

### Exercise 3

$$(i) \text{ Calculate } \begin{pmatrix} 1 & 2 \\ -1 & 3 \end{pmatrix} * \begin{pmatrix} 3 & -1 \\ 2 & 1 \end{pmatrix}$$

$$(ii) \text{ Calculate } \begin{pmatrix} 1 & 3 & 5 \\ -1 & 2 & -1 \end{pmatrix} * \begin{pmatrix} 1 & 1 \\ 2 & 2 \\ 0 & -1 \end{pmatrix}$$

- (iii) Is  $*$  (a) commutative?  
(b) associative?

Besides the composition of vector space morphisms, we can also define the following:

If  $T_1$  and  $T_2$  are morphism mapping a vector space  $V$  to a vector space  $U$ ,



then we have two vector space operations defined on  $U$  which can be used to define new morphism of  $V$  to  $U$ :

$$T_1 + T_2 : \underline{v} \longmapsto T_1(\underline{v}) + T_2(\underline{v}) \quad (\underline{v} \in V)$$

$$\alpha T_1 : \underline{v} \longmapsto \alpha(T_1(\underline{v})) \quad (\underline{v} \in V)$$

where  $\alpha$  is a real number.

The verification of this result is not difficult and follows the same lines as the proof above for composition of morphisms. What interests us here are the corresponding matrix operations. It is fairly easy to see that these are as follows:

- (i) If  $k$  is any real number, then  $kA$  is the matrix obtained by multiplying each element of  $A$  by  $k$ .
- (ii) If  $A$  and  $B$  are two matrices each with  $n$  rows and  $m$  columns, then  $A \triangle B$  is the matrix with  $n$  rows and  $m$  columns obtained by adding corresponding elements of  $A$  and  $B$ .

#### Example 4

- (i)  $3 \begin{pmatrix} 1 & 2 \\ -1 & 3 \\ 4 & 7 \end{pmatrix} = \begin{pmatrix} 3 & 6 \\ -3 & 9 \\ 12 & 21 \end{pmatrix}$
- (ii)  $\begin{pmatrix} 1 & 2 \\ -1 & 3 \\ 4 & 7 \end{pmatrix} \triangle \begin{pmatrix} 3 & 6 \\ -3 & 9 \\ 12 & 21 \end{pmatrix} = \begin{pmatrix} 1+3 & 2+6 \\ (-1)+(-3) & 3+9 \\ 4+12 & 7+21 \end{pmatrix}$   
 $= \begin{pmatrix} 4 & 8 \\ -4 & 12 \\ 16 & 28 \end{pmatrix}$
- (iii)  $\begin{pmatrix} 1 & 2 \\ -1 & 3 \end{pmatrix} \triangle \begin{pmatrix} 3 & -3 & 12 \\ 6 & 9 & 21 \end{pmatrix}$

is not defined: it is equivalent to trying to “add” two functions with different domains.

It is clear from the definition of  $\triangle$  that it is both commutative and



associative. Because of its obvious similarities to addition, we use the symbol  $+$  for  $\triangle$ ; for example, we write

$$\begin{pmatrix} 2 & 3 \\ 1 & 4 \end{pmatrix} + \begin{pmatrix} 2 & 7 \\ -2 & -4 \end{pmatrix} = \begin{pmatrix} 4 & 10 \\ -1 & 0 \end{pmatrix}$$

and we call the operation **matrix addition**.

We can use these two operations to define a third operation—**matrix subtraction**—by

$$A - B = A + (-1)B$$

The matrix  $(-1)B$  is usually written as  $-B$ .

So now we have three operations:  $*$ , multiplication by a scalar, and  $+$ . The two operations  $+$  and  $*$  are binary operations on sets of matrices although we have to take some care over defining the sets in which they operate. We have noted some of their properties:  $+$  is commutative and associative;  $*$  is associative but not commutative (see Exercise 3). We have not as yet seen how  $*$  and  $+$  interact. In particular, are they distributive over each other?

#### Exercise 4

By choosing some simple matrices, prove that  $+$  is not distributive over  $*$ , i.e. that for some suitable† matrices  $A$ ,  $B$  and  $C$

$$A + (B * C) \neq (A + B) * (A + C)$$

It is a relatively simple matter to disprove a conjecture by a counter-example, as in Exercise 4, but to *prove* a conjecture we have to argue in general terms. We shall not do this here, because the proof does not illustrate anything very important (except, perhaps, a need for an improved notation, which will be developed later); however, we record the fact that  $*$  *is* distributive over  $+$ . That is, for suitable matrices  $A$ ,  $B$  and  $C$ ,

$$A * (B + C) = (A * B) + (A * C)$$

and

$$(A + B) * C = (A * C) + (B * C)$$

Notice that we must make *both* statements because  $*$  is not commutative.

† By “suitable” we mean matrices with appropriate numbers of rows and columns, so that the operations are defined.



There is one difficulty that we must mention; it concerns terminology. It is widespread practice to refer to the operation  $*$  as *matrix multiplication* and drop the symbol  $*$ , writing  $AB$  instead of  $A * B$ . This is most unfortunate, for although  $*$  is associative, and distributive over  $+$ , just as for  $\times$  and  $+$  in the set of real numbers,  $*$  is *not* commutative, and whereas  $+$  comes from addition of mappings,  $*$  comes not from multiplication but from composition.

However, because the practice is widespread we shall refer to the operation as *matrix multiplication*, and drop the symbol  $*$  after the next section.

## 6.4 Some Special Matrices

### Rotation Matrices

In Example 5.4.3 we considered a morphism of  $R^2$  to  $R^2$  which, in terms of equations, is specified by

$$x'_1 = -x_2 = 0x_1 + (-1)x_2$$

$$x'_2 = x_1 = 1x_1 + 0x_2$$

We saw that this can be regarded as specifying a mapping which rotates

the plane through an angle  $\frac{\pi}{2}$  anti-clockwise about the origin. In terms of

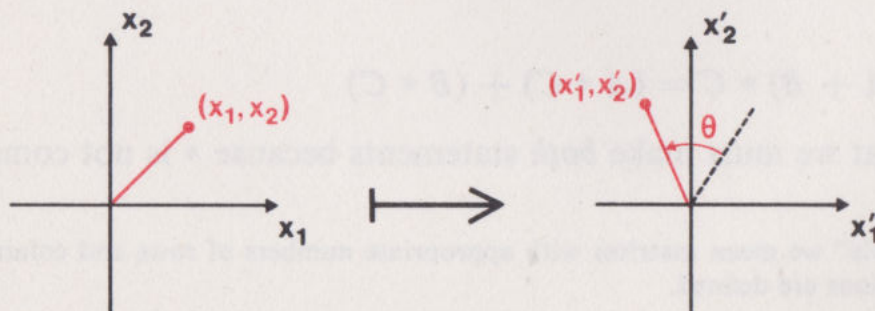
matrices, we can say that  $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$  represents a rotation of the plane

through  $\frac{\pi}{2}$  anti-clockwise.

Suppose we start the other way round and look for matrices to do some specified geometric job.

#### Example 1

Consider a mapping which rotates the plane about the origin through an angle  $\theta$  anti-clockwise.





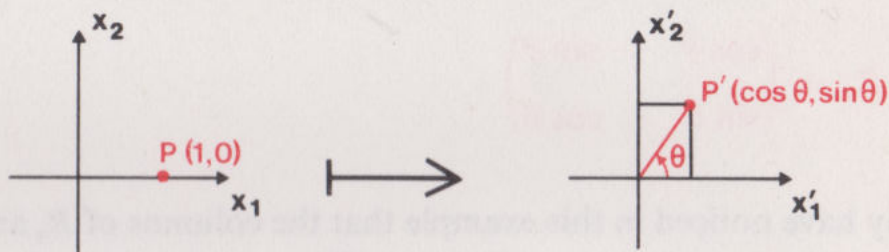
If we call the corresponding matrix  $R_\theta$ , then

$$\begin{pmatrix} x'_1 \\ x'_2 \end{pmatrix} = R_\theta \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

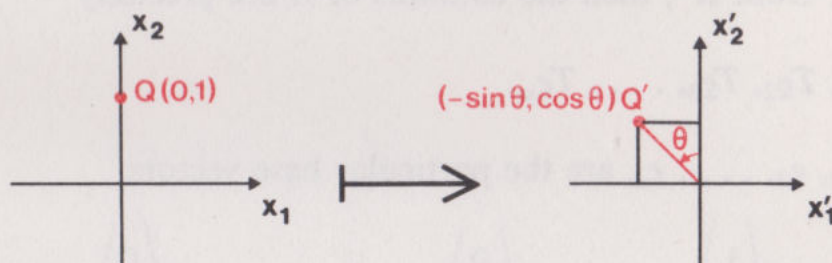
where  $x_1, x_2, x'_1$  and  $x'_2$  are related as in the diagram. We can find  $R_\theta$  easily by considering the basis  $\left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\}$  for  $R^2$ .

A little trigonometry tells us that

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} \mapsto \begin{pmatrix} \cos \theta \\ \sin \theta \end{pmatrix}$$



$$\text{and } \begin{pmatrix} 0 \\ 1 \end{pmatrix} \mapsto \begin{pmatrix} -\sin \theta \\ \cos \theta \end{pmatrix}$$



$R_\theta$  has two rows and two columns because the mapping is one-one and the image set of  $R^2$  is  $R^2$ . Thus if

$$R_\theta = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

we know that

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} * \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \cos \theta \\ \sin \theta \end{pmatrix}$$

and

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} * \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -\sin \theta \\ \cos \theta \end{pmatrix}$$



Since

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} * \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} a \\ c \end{pmatrix}$$

and

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} * \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} b \\ d \end{pmatrix}$$

we find that

$$a = \cos \theta, \quad c = \sin \theta, \quad b = -\sin \theta \text{ and } d = \cos \theta.$$

Thus

$$R_\theta = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}$$

You may have noticed in this example that the columns of  $R_\theta$  are precisely the images of the base vectors  $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$  and  $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$  under the mapping. This is not special to this example, for it is easily seen that if  $A$  is the matrix of a morphism  $T$  from  $R^m$ , then the columns of  $A$  are precisely

$$Te_1, Te_2, Te_3, \dots, Te_m,$$

where  $e_1, e_2, e_3, \dots, e_m$  are the particular base vectors

$$e_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad e_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \dots, \quad e_m = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$$

### Exercise 1

Find the matrix which corresponds to a “stretching of the plane”, so that the distance of every point from the origin is multiplied by a number  $k$ .



*Exercise 2*

Find the matrix which corresponds to the identity mapping, which leaves every point in the plane unchanged.

**The Identity Matrix**

Under the operation of composition of functions, the identity function leaves every function unchanged. For example, the function:

$$f: x \longmapsto x \quad (x \in R)$$

is such that  $f \circ g = g \circ f = g$  for any function  $g$  with domain and codomain  $R$ . This property has its counterpart in matrix algebra. If we write

$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ , then  $I * A = A * I = A$  for *any* matrix with two rows and

two columns. Indeed, if we write  $I_{m,m}$  for the matrix with  $m$  rows and  $m$  columns, such that every entry is zero except for 1's as the diagonal entries from the top left-hand corner to the bottom right-hand corner, then if  $A$  is any matrix with  $m$  rows and  $n$  columns, we have

$$I_{m,m} * A = A * I_{n,n} = A$$

Usually, the suffices are dropped and we just write  $I$  for the identity matrix appropriate to the particular situation.

In the next section we collect together some of the differences between matrices and numbers, and also some of the similarities of the two algebras.

**6.5 Matrix Algebra and the Algebra of Numbers**

A major difference between matrices and numbers is that, whereas we can add or multiply any two numbers, this is not the case for matrices. If we restrict ourselves to a set of matrices with the same (fixed) number of rows as columns, then we can add or multiply any two matrices from the set. We take it as implicit, then, in this section that *all the matrices have the same number of rows as columns, and that this number is the same for all the matrices*. (A matrix with the same number of rows as columns is called a *square matrix*.)



### Similarities

	Numbers	Matrices
Addition is commutative.	$a + b = b + a$	$A + B = B + A$
Addition is associative.	$(a + b) + c = a + (b + c)$	$(A + B) + C = A + (B + C)$
There is a zero element.	$a + 0 = 0 + a = a$	$A + O = O + A = A$ , where $O$ is the matrix in which every entry is zero.
Each element has an additive inverse.	$(-b) + b = 0$	$(-B) + B = O$
Multiplication is associative.	$(ab)c = a(bc)$	$(AB)C = A(BC)$
There is an "identity" element.	$a1 = 1a = a$	$AI = IA = A$ , where $I$ is the identity matrix defined on page 221
"Multiplication" is distributive over addition.	$a(b + c) = ab + ac$ $(a + b)c = ac + bc$	$A(B + C) = AB + AC$ $(A + B)C = AC + BC$
Certain product rules hold.	$a0 = 0a = 0$ $a(-b) = -ab$ $(-a)(-b) = ab$	$AO = OA = O$ $A(-B) = -AB$ $(-A)(-B) = AB$

### Differences

- (i) Number multiplication is commutative; matrix multiplication is not. That is, there exist matrices  $A, B$  such that  $AB \neq BA$ .

As a consequence, the matrix expansion  $(A + B)^2 = A^2 + 2AB + B^2$  is *false*, in general. The correct expansion is  $(A + B)^2 = A^2 + AB + BA + B^2$ . The matrix  $AB + BA$  is equal to  $2AB$  if and only if  $A$  and  $B$  commute, i.e.  $AB = BA$ . This may happen in certain cases (for example if  $B = I$ ), but it is not generally true.

- (ii) The product of two non-zero numbers is non-zero, but the product of two non-zero matrices may be equal to the zero matrix.

For example, if

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \text{ and } B = \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix},$$

then  $AB = O$ . (Note also that  $BA \neq O$ .) This example also illustrates the existence of a non-zero matrix  $B$  such that  $B^2 = O$ .

- (iii) The cancellation law holds for numbers, but not for matrices. That is, if  $a, b, c$  are numbers and  $a \neq 0$ , then

$$(ab = ac) \text{ implies } (b = c)$$

But for matrices,

$$(AB = AC) \text{ does not imply } (B = C).$$



For example, if  $A$  and  $B$  are as above, then  $AB = AO$ , but  $B$  is not equal to the zero matrix, i.e.  $A$  cannot be cancelled.

- (iv) The equation  $x^2 = 0$  has the unique solution  $x = 0$ , but the matrix equation  $X^2 = O$  has an infinite number of solutions.

It can be shown that the general solution of the matrix equation  $X^2 = O$  for matrices with two rows and columns is

$$X = \begin{pmatrix} r & s \\ -\frac{r^2}{s} & -r \end{pmatrix} \text{ for arbitrary } r \text{ and non-zero } s,$$

or

$$X = \begin{pmatrix} 0 & 0 \\ t & 0 \end{pmatrix} \text{ for arbitrary } t.$$

- (v) The equation  $x^2 = -1$  has no solution (in real numbers), but the matrix equation  $X^2 = -I$  does have solutions.

We single out two solutions for special attention. These are the matrices

$$C = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \text{ and } -C = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

Note that the matrix  $-I$  is equal to the rotation matrix  $R_\pi$  corresponding to a rotation about the origin through the angle  $\pi$ . The matrix  $C$  is the rotation matrix  $R_{\frac{1}{2}\pi}$ . What transformation does  $-C$  correspond to?

### Exercise 1

Give expansions of  $(A - B)^2$  and  $(A - B)^3$  which are valid for all square matrices. (There are 8 terms in the second expansion.)

### Exercise 2

The equation  $x^2 = x$  has precisely two solutions,  $x = 0$  or  $1$ . By giving an example of a matrix with 2 rows and 2 columns, show that the matrix equation  $X^2 = X$  has solutions other than  $X = O$  or  $I$ .



## 6.6 Additional Exercises

### Exercise 1

- (i) What is the result of rotating the plane anti-clockwise about the origin through an angle  $\alpha$  and then through an angle  $\beta$ ?
- (ii) Complete the following equation:

$$R_\beta * R_\alpha \boxed{\phantom{0000}}$$

- (iii) Complete the following equation

$$R_{\beta+\alpha} = \begin{pmatrix} \cos(\beta + \alpha) & \boxed{\phantom{0000}} \\ \boxed{\phantom{0000}} & \boxed{\phantom{0000}} \end{pmatrix}$$

- (iv) Write down  $R_\alpha$  and  $R_\beta$  and calculate  $R_\beta * R_\alpha$  directly.
- (v) Compare your answers to (iii) and (iv). Can you deduce anything from this comparison?

### Exercise 2

Find two non-zero matrices with real elements,  $X$ , which satisfy the equation

$$\begin{pmatrix} 1 & 2 \\ 0 & 0 \end{pmatrix} X = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

### Exercise 3

The lines  $L_1$ ,  $L_2$  and  $L_3$  have the equations

$$y = 0$$

$$x = 0$$

$$x = -y$$

respectively.

Determine in matrix form all the morphisms of  $R^2$  to  $R^2$  which

- (i) map  $L_1$  to  $L_1$
- (ii) map  $L_2$  to  $L_2$
- (iii) map  $L_3$  to  $L_3$ .



*Exercise 4*

The set of all  $2 \times 2$  matrices with real elements under the operation of addition forms a vector space. Determine the dimension of this vector space and specify any two different sets of vectors each of which forms a basis.

**6.7 Answers to Exercises****Section 6.1***Exercise 1*

Equations (1) do not define a morphism because  $\sin(\alpha + \beta) \neq \sin \alpha + \sin \beta$  in general.

Equations (2) do define a morphism.

**Section 6.3***Exercise 1*

- (i)  $\begin{pmatrix} 5 \\ 6 \\ 1 \end{pmatrix}$ , because  $1 \times 3 + (-1) \times (-2) = 5$   
 $2 \times 3 + 0 \times (-2) = 6$   
 $3 \times 3 + 4 \times (-2) = 1$
- (ii)  $\begin{pmatrix} 14 \\ 11 \end{pmatrix}$ , because  $1 \times 1 + 2 \times 2 + 3 + 3 = 14$   
 $(-1) \times 1 + 0 \times 2 + 4 + 3 = 11$

*Exercise 2*

Substituting from (i) into (ii), we get

$$x_1'' = e(ax_1 + bx_2) + f(cx_1 + dx_2)$$

$$x_2'' = g(ax_1 + bx_2) + h(cx_1 + dx_2)$$

i.e.

$$x_1'' = (ea + fc)x_1 + (eb + fd)x_2$$

$$x_2'' = (ga + hc)x_1 + (gb + hd)x_2$$

and the matrix representing these equations is

$$\begin{pmatrix} ea + fc & eb + fd \\ ga + hc & gb + hd \end{pmatrix}$$



## Exercise 3

(i)  $\begin{pmatrix} 7 & 1 \\ 3 & 4 \end{pmatrix}$

(ii)  $\begin{pmatrix} 7 & 2 \\ 3 & 4 \end{pmatrix}$

- (iii) (a) It is easy enough to find a counter-example to show that
- $*$
- is not commutative. For example,

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} * \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

and

$$\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} * \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$$

- (b) This is more difficult, because
- $*$
- is in fact, associative.

A direct proof for  $*$  could be tricky, but we can go about it another way. Remembering that  $*$  for matrices corresponds to composition of morphisms, which are just special kinds of functions. we can show that the composition of functions (where it is possible) is associative. Suppose  $A, B, C, D$  are sets and  $f, g, h$  are functions such that

$$f: A \longrightarrow B, g: B \longrightarrow C, h: C \longrightarrow D$$

then if  $a \in A$ ,

$$\begin{aligned}
 ((h \circ g) \circ f)(a) &= (h \circ g)(f(a)) \\
 &= h(g(f(a))) \\
 &= h((g \circ f)(a)) \\
 &= (h \circ (g \circ f))(a)
 \end{aligned}$$

Thus, composition of functions is associative, which implies that  $*$  is also associative.

## Exercise 4

For example, with

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}, B = \begin{pmatrix} 1 & -2 \\ 3 & -1 \end{pmatrix}, C = \begin{pmatrix} 1 & 1 \\ 3 & -2 \end{pmatrix},$$



we have

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} + \left[ \begin{pmatrix} 1 & -2 \\ 3 & -1 \end{pmatrix} * \begin{pmatrix} 1 & 1 \\ 3 & -2 \end{pmatrix} \right] = \begin{pmatrix} -4 & 7 \\ 3 & 9 \end{pmatrix}$$

whereas

$$\left[ \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} + \begin{pmatrix} 1 & -2 \\ 3 & -1 \end{pmatrix} \right] * \left[ \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} + \begin{pmatrix} 1 & 1 \\ 3 & -2 \end{pmatrix} \right] = \begin{pmatrix} 4 & 6 \\ 30 & 24 \end{pmatrix}$$

## Section 6.4

### Exercise 1

Since  $\begin{pmatrix} 1 \\ 0 \end{pmatrix} \mapsto \begin{pmatrix} k \\ 0 \end{pmatrix}$  and  $\begin{pmatrix} 0 \\ 1 \end{pmatrix} \mapsto \begin{pmatrix} 0 \\ k \end{pmatrix}$ , the matrix required is  $\begin{pmatrix} k & 0 \\ 0 & k \end{pmatrix}$ .

### Exercise 2

The matrix is  $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ .

(The special case of Exercise 1 when  $k = 1$ .)

## Section 6.5

### Exercise 1

$$(A - B)^2 = A^2 - AB - BA + B^2$$

$$(A - B)^3 = A^3 - A^2B - ABA + AB^2 - BA^2 + BAB + B^2A - B^3$$

### Exercise 2

$$X = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \text{ is an example.}$$



## Section 6.6

## Exercise 1

If we perform two rotations such as these successively, the result is a rotation through an angle  $\beta + \alpha$ , which is of course the same as a rotation through  $\alpha + \beta$ . Hence

(i) A rotation through  $\beta + \alpha$ .

(ii)  $R_\beta * R_\alpha = R_{\beta+\alpha}$

$$(iii) R_{\beta+\alpha} = \begin{pmatrix} \cos(\beta + \alpha) & -\sin(\beta + \alpha) \\ \sin(\beta + \alpha) & \cos(\beta + \alpha) \end{pmatrix}$$

$$(iv) R_\beta * R_\alpha = \begin{pmatrix} \cos \beta & -\sin \beta \\ \sin \beta & \cos \beta \end{pmatrix} * \begin{pmatrix} \cos \alpha & -\sin \alpha \\ \sin \alpha & \cos \alpha \end{pmatrix} \\ = \begin{pmatrix} \cos \beta \cos \alpha - \sin \beta \sin \alpha & -\cos \beta \sin \alpha - \sin \beta \cos \alpha \\ \sin \beta \cos \alpha + \cos \beta \sin \alpha & -\sin \beta \sin \alpha + \cos \beta \cos \alpha \end{pmatrix}$$

(v) Comparing (iii) and (iv), we obtain the formulas

$$\cos(\beta + \alpha) = \cos \beta \cos \alpha - \sin \beta \sin \alpha$$

$$\sin(\beta + \alpha) = \sin \beta \cos \alpha + \cos \beta \sin \alpha$$

## Exercise 2

Let

$$X = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad a, b, c, d \in R.$$

Then we require

$$a + 2c = 0 \quad \text{i.e. } a = -2c$$

$$b + 2d = 0 \quad \text{i.e. } b = -2d$$

So any matrix of the form  $\begin{pmatrix} -2c & -2d \\ c & d \end{pmatrix}$ , in which not both  $c$  and  $d$  are zero, can be  $X$ .

For example:  $\begin{pmatrix} 2 & 2 \\ -1 & -1 \end{pmatrix}$  and  $\begin{pmatrix} -4 & -4 \\ 2 & 2 \end{pmatrix}$

## Exercise 3

(i) The basis vector  $(1, 0)$  must map to itself or a multiple of itself, viz.:  $(a, 0)$ ,  $a \in R$ . The image of the other basis vector  $(0, 1)$  is immaterial: let  $(0, 1)$  map to  $(c, d)$ ,  $c, d \in R$ .



Then the morphism has matrix

$$\begin{pmatrix} a & c \\ 0 & d \end{pmatrix}$$

(ii) By a similar argument to 1 above, the morphism has matrix

$$\begin{pmatrix} a & 0 \\ b & d \end{pmatrix}, a, b, d \in R.$$

(iii) If  $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$  is the matrix of the morphism, then since any vector  $(x, -x)$  must map to one in the same direction, i.e.  $(kx, -kx)$ ,  $k \in R$ ,  $k \neq 0$ , we have

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ -x \end{pmatrix} = \begin{pmatrix} kx \\ -kx \end{pmatrix}$$

i.e.

$$a - b = k$$

$$c - d = -k$$

Thus any matrix of the form  $\begin{pmatrix} b+k & b \\ c & c+k \end{pmatrix}$  will map  $L_3$  to  $L_3$ .

#### Exercise 4

The dimension of the vector space of all  $2 \times 2$  real matrices under addition has dimension 4, since a  $2 \times 2$  matrix has 4 elements.

Any set of four linearly independent matrices forms a basis.

Examples of such sets are

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \quad \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \quad \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \quad \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

or  $\begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix} \quad \begin{pmatrix} 0 & 2 \\ 0 & 0 \end{pmatrix} \quad \begin{pmatrix} 0 & 0 \\ 2 & 0 \end{pmatrix} \quad \begin{pmatrix} 0 & 0 \\ 0 & 2 \end{pmatrix}$

or  $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \quad \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \quad \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \quad \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$  etc.



# CHAPTER 7 LINEAR EQUATIONS AND MATRICES

## 7.0 Introduction

In this chapter we consider some theoretical aspects of solving systems of simultaneous linear equations. We use the ideas of the previous chapters in an investigation of the existence and uniqueness of solutions to such systems of equations. This chapter prepares the ground for the next chapter in which practical methods of solving systems of simultaneous equations will be discussed.

A continuing theme of this chapter is the notion of a *morphism* (also known, in the context of vector spaces, as a *linear transformation*) between two vector spaces. In particular, we shall discuss under what circumstances an inverse morphism exists.

We have already considered the idea of simultaneous equations and how they relate to mappings between vector spaces. Before we discuss this in greater detail, we shall explain what is meant by a *system of simultaneous linear equations* and a *solution set* of such a system. This will lead us to consider how to find a solution and whether the solution is unique—and indeed whether a solution exists at all.

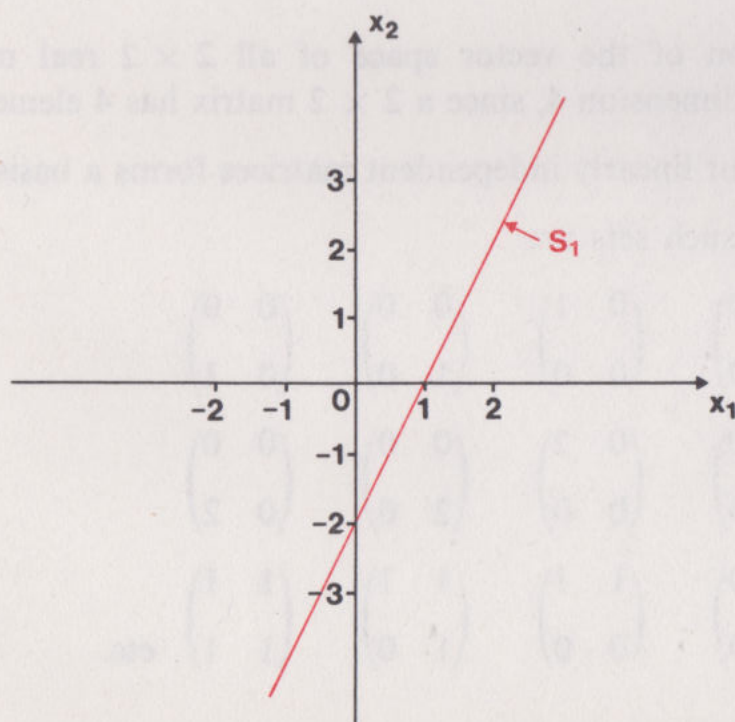
We can consider the equation

$$2x_1 - x_2 = 2,$$

as defining the set of all ordered pairs of real numbers  $(x_1, x_2)$  such that  $2x_1 - x_2 = 2$ ; that is,

$$S_1 = \{(x_1, x_2) : 2x_1 - x_2 = 2\},$$

which can be represented diagrammatically as in the following figure:

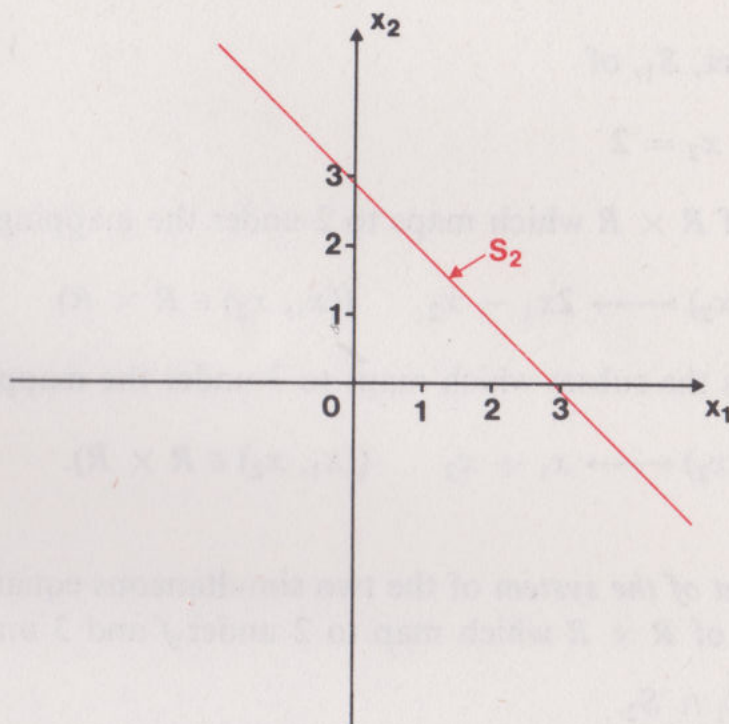




Similarly, if we consider the equation  $x_1 + x_2 = 3$ , it defines the set  $S_2$ , where

$$S_2 = \{(x_1, x_2): x_1 + x_2 = 3\}.$$

$S_2$  is represented in the figure below:



We now have two equations:

$$2x_1 - x_2 = 2$$

$$x_1 + x_2 = 3.$$

If we consider these two equations together, they form a *system of simultaneous equations* in the sense that we consider just those pairs  $(x_1, x_2)$  which satisfy both equations. Each of the equations has two variables,  $x_1$  and  $x_2$ , and each equation defines a straight line in the plane. We therefore say that the equations are *linear*. Thus the system of equations is a set of two simultaneous linear equations in two variables,  $x_1$  and  $x_2$ :

In this chapter we shall deal with systems of linear equations, but we shall generalize to consider  $m$  simultaneous equations in  $n$  variables (also called *unknowns*). A **linear equation** in  $n$  variables,  $x_1, x_2, \dots, x_n$ , is an equation of the form

$$a_1x_1 + a_2x_2 + \dots + a_nx_n = b,$$

where  $a_1, a_2, \dots, a_n, b$  are known numbers. As we have seen, for two variables we can represent such an equation by a line; for three variables



the representation is a plane; for more than three variables we have no visual representation, but geometers say that such an equation represents a *hyper-plane*.

We now consider the solution set of a system of simultaneous linear equations.

The solution set,  $S_1$ , of

$$2x_1 - x_2 = 2$$

is the subset of  $R \times R$  which maps to 2 under the mapping

$$f:(x_1, x_2) \longmapsto 2x_1 - x_2 \quad ((x_1, x_2) \in R \times R).$$

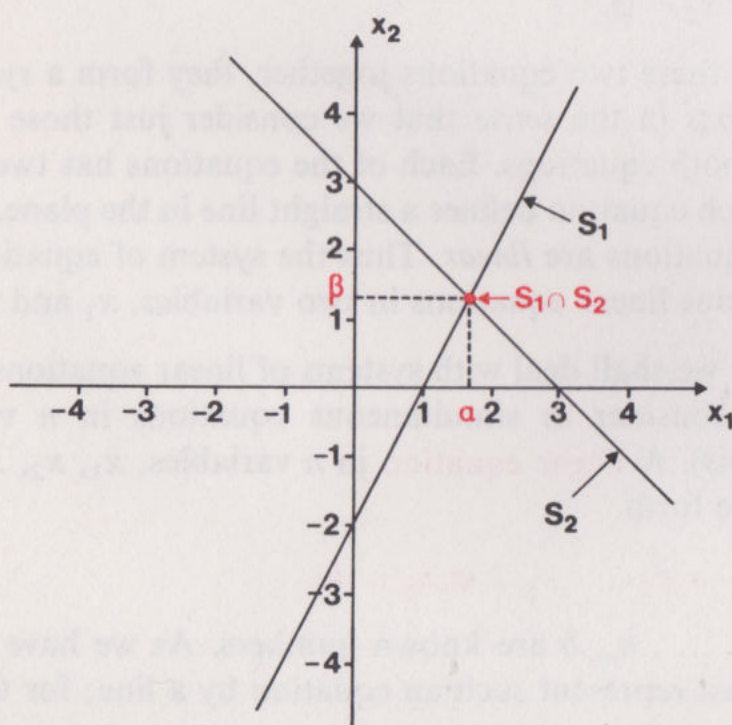
Similarly,  $S_2$  is the subset which maps to 3 under the mapping

$$g:(x_1, x_2) \longmapsto x_1 + x_2 \quad ((x_1, x_2) \in R \times R).$$

The *solution set of the system* of the two simultaneous equations is the set  $S$  of elements of  $R \times R$  which map to 2 under  $f$  and 3 under  $g$ ; that is,

$$S = S_1 \cap S_2.$$

In other words, the solution set of the system is the set of ordered pairs  $\{(x_1, x_2)\}$  such that each element of this set belongs *both* to the set  $S_1$  and to the set  $S_2$ .





We can see from the above diagram that  $S$  consists of one element only, namely the point of intersection of the two straight lines. If the co-ordinates of this point are  $\alpha$  and  $\beta$ , then

$$S = \{(\alpha, \beta)\}.$$

For a system of  $m$  simultaneous linear equations in  $n$  variables,  $x_1, x_2, \dots, x_n$ , the **solution set**  $S$  is the subset of  $R^n$  defined by

$$S = S_1 \cap S_2 \cap \dots \cap S_m,$$

where  $S_i$  is the solution set of the  $i$ th equation (a set of  $n$ -tuples),  $i = 1, 2, \dots, m$ .

## 7.1 The Nature of the Solution I

Consider the solution set of the system of equations:

$$-2x_1 + 3x_2 = 6$$

$$2x_1 - 3x_2 = 12$$

The graphs of the equations are two parallel lines and hence have no common point, so that if

$$S_1 = \{(x_1, x_2) : -2x_1 + 3x_2 = 6\},$$

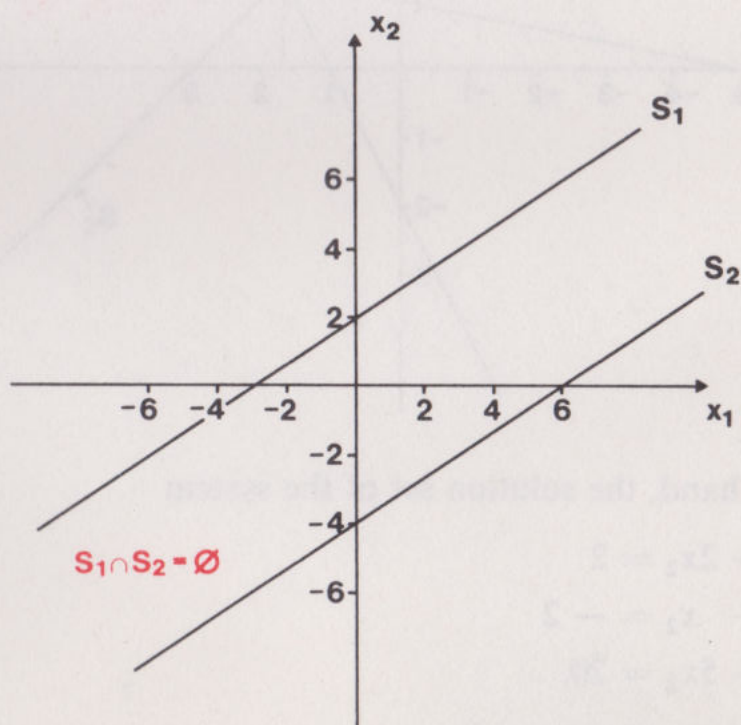
and

$$S_2 = \{(x_1, x_2) : 2x_1 - 3x_2 = 12\},$$

then

$$S = S_1 \cap S_2 = \emptyset.$$

The solution set is empty, and we say that the system of equations has *no solution*.





There are examples of systems of linear equations for which the corresponding graphs are non-parallel straight lines and yet the solution sets are empty.

If we consider the equations

$$2x_1 - x_2 = 2$$

$$x_1 + x_2 = 3$$

$$-x_1 + 5x_2 = 5,$$

they define the sets

$$S_1 = \{(x_1, x_2) : 2x_1 - x_2 = 2\}$$

$$S_2 = \{(x_1, x_2) : x_1 + x_2 = 3\}$$

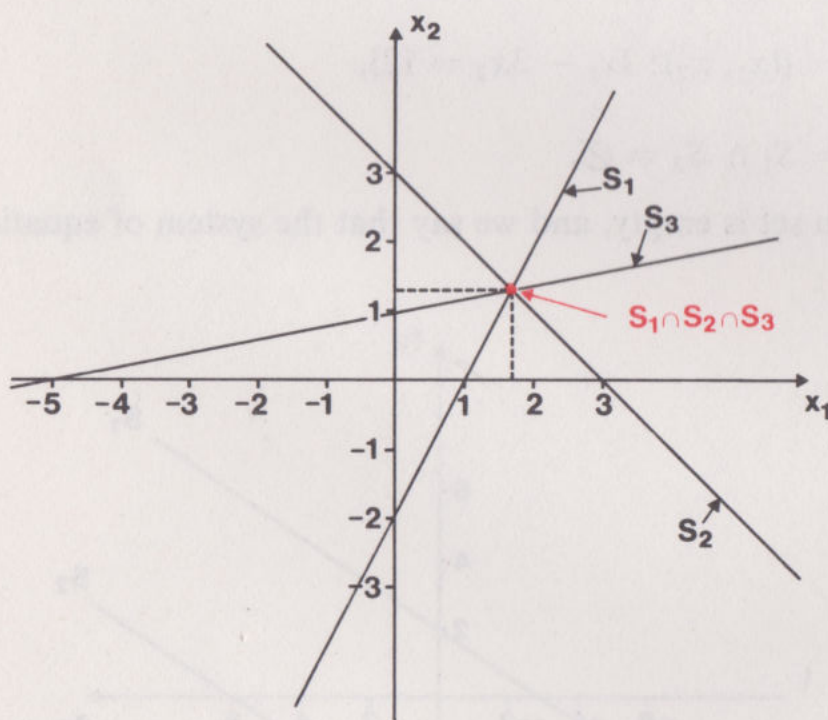
$$S_3 = \{(x_1, x_2) : -x_1 + 5x_2 = 5\},$$

and the solution set  $S$  of the system is

$$S = S_1 \cap S_2 \cap S_3$$

$$= \left\{ \left( \frac{5}{3}, \frac{4}{3} \right) \right\}.$$

This system has only *one* solution.



On the other hand, the solution set of the system

$$x_1 - 2x_2 = 2$$

$$x_1 - x_2 = -2$$

$$4x_1 + 5x_2 = 20,$$



which defines the three sets

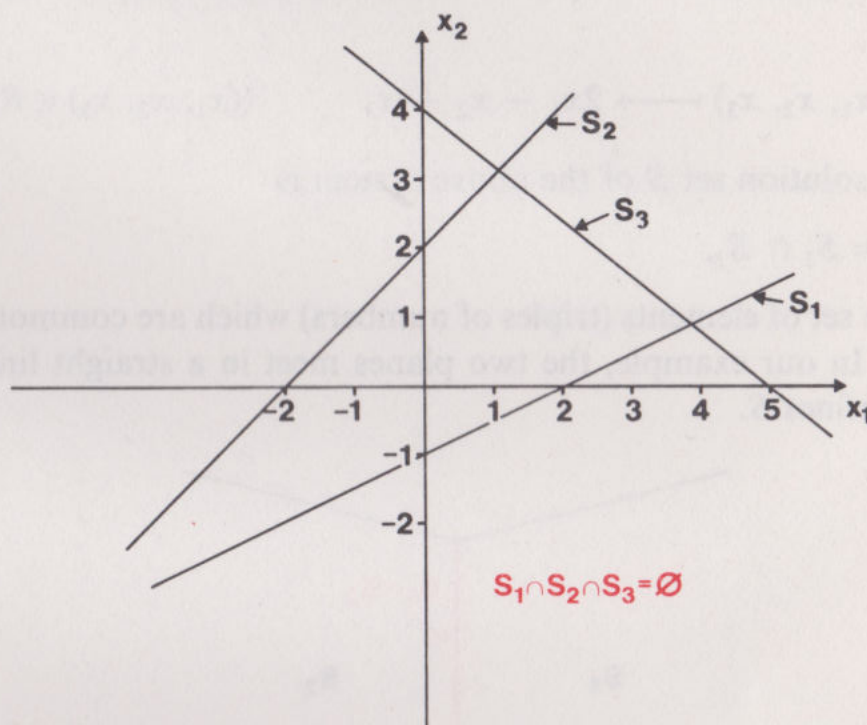
$$S_1 = \{(x_1, x_2): x_1 - 2x_2 = 2\}$$

$$S_2 = \{(x_1, x_2): x_1 - x_2 = -2\}$$

$$S_3 = \{(x_1, x_2): 4x_1 + 5x_2 = 20\},$$

is empty; that is,

$$S = S_1 \cap S_2 \cap S_3 = \emptyset.$$



The figure shows that although any two of the lines intersect, so that

$$S_1 \cap S_2 \neq \emptyset, S_1 \cap S_3 \neq \emptyset \text{ and } S_2 \cap S_3 \neq \emptyset,$$

nevertheless, the three lines do not all intersect at a common point, so that

$$S_1 \cap S_2 \cap S_3 = \emptyset.$$

It follows that the system of equations has *no solution*.

An equation of the type

$$ax_1 + bx_2 + cx_3 = d$$

(where  $a, b, c$  and  $d$  are constants) can be represented by a plane in a three-dimensional geometric space. The two equations

$$2x_1 - x_2 + x_3 = 4$$

$$x_1 + 3x_2 - 2x_3 = 9$$



define two sets,  $S_1$  and  $S_2$ , of ordered triples:

$$S_1 = \{(x_1, x_2, x_3): 2x_1 - x_2 + x_3 = 4\}$$

$$S_2 = \{(x_1, x_2, x_3): x_1 + 3x_2 - 2x_3 = 9\}.$$

The solution set of an equation involving three variables, for example,

$$2x_1 - x_2 + x_3 = 4$$

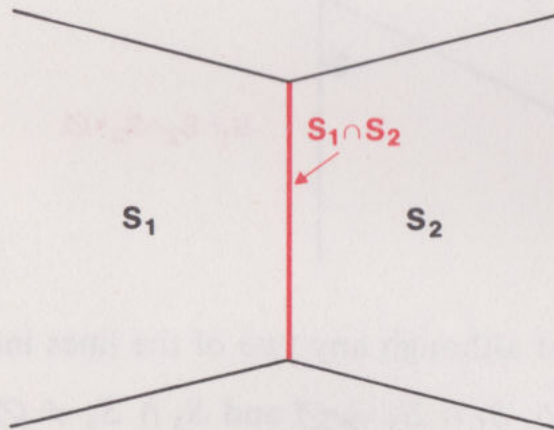
is the set of elements of  $R^3$  (i.e.  $R \times R \times R$ ) which map to 4 under the mapping

$$f: (x_1, x_2, x_3) \longmapsto 2x_1 - x_2 + x_3 \quad ((x_1, x_2, x_3) \in R^3).$$

Again, the solution set  $S$  of the above system is

$$S = S_1 \cap S_2,$$

which is the set of elements (triples of numbers) which are common to both  $S_1$  and  $S_2$ . In our example, the two planes meet in a straight line, which therefore defines  $S$ .



In this case, the system of equations has *more than one solution*.

On the other hand, the two planes defined by the equations

$$2x_1 - x_2 - x_3 = 4$$

$$4x_1 - 2x_2 - 2x_3 = 7$$

are parallel. The planes do not intersect and the set  $S$  is empty, so this system has *no solution*.

In general, for any system of linear equations, we distinguish three types of solution set:

- (i) a solution set which is empty, in which case we say that the system has **no solution**;



- (ii) a solution set which contains just one element (an  $n$ -tuple, for a system of equations in  $n$  variables), in which case we say that the system has a **unique solution**;
- (iii) a solution set which contains more than one element, in which case we say that the system has **more than one solution**.

### Exercise 1

Examine each of the following systems of linear equations. Distinguish the three types:

- A: no solution;  
 B: unique solution;  
 C: more than one solution.

Indicate, by a tick in the appropriate box, the class to which each system belongs. (There is no need to solve the equations: a geometrical argument is sufficient, but in case you find it easier to use an algebraic argument, we give some details in the solution.)

	A	B	C
(i) $x_1 - x_2 = 4$ $2x_1 + x_2 = 3$			
(ii) $x_1 + 2x_2 = 1$ $-x_1 + 3x_2 = 0$ $x_1 + x_2 = 2$			
(iii) $3x_1 - 4x_2 = -1$ $-3x_1 + 4x_2 = 1$			
(iv) $-x_1 + x_2 = 2$ $x_2 + x_3 = 0$ $x_1 + x_3 = 1$			

### Note

Notice that we ought really to specify how many variables are involved. Thus in (iv), three variables are involved, although each individual equation involves only two. We could make this clear by writing

$$-x_1 + x_2 + 0x_3 = 2,$$

and so on. (Where these equations arise in practice the number of variables



is usually clear from the context.) There is a considerable difference between the system in (i) which is taken to involve two variables, and the system

$$x_1 - x_2 + 0x_3 = 4$$

$$2x_1 + x_2 + 0x_3 = 3,$$

so we specify that in the first three parts of this exercise, two variables are involved.

## 7.2 Solving Systems of Linear Equations

It may seem fairly obvious to you that the solution set of a system of simultaneous equations is unaffected by the following **elementary operations**:

- (i) interchanging any two equations of the system;
- (ii) multiplying every term in an equation by some non-zero constant;
- (iii) adding a multiple of one equation to another equation.

We could give some justification for the assertion that these operations do not change the solution set of a system, but without an axiomatic basis for our discussion such an argument would have no real validity.

### The Method of Gauss Elimination

Although the three elementary operations do not change the solution set, they do change the *equations* of the system considered. We obtain a different system of equations which has the same solution set as the original system. Any two linear systems having the **same solution set** are said to be **equivalent systems**.

Many methods of solving systems of simultaneous linear equations depend on finding an equivalent system for which it is simple to find the solution set. We shall illustrate this by solving the system of equations:

$$2x_1 + x_2 - x_3 = 3$$

$$6x_1 - x_2 - 9x_3 = 7$$

$$4x_1 + 3x_2 + x_3 = 5$$

We shall use a method known as the **Gauss elimination method** to solve this system; the method is a formalization of a procedure with which you are probably familiar. We begin by adding appropriate multiples of the first equation to each of the other two so as to eliminate  $x_1$  from the latter two equations. We then add an appropriate multiple of the (new) second



equation of the (new) third equation so as to eliminate  $x_2$  from the latter. We then have a system of equations, equivalent to the original system, which can be solved easily.

If we denote the three equations in each of the equivalent systems by  $R_1$ ,  $R_2$  and  $R_3$  in that order, then we can outline the above sequence of operations in the following diagram.

Action	System	
	System of equations	Variables
	$R_1$	$x_1, x_2, x_3$
	$R_2$	$x_1, x_2, x_3$
	$R_3$	$x_1, x_2, x_3$
Eliminate variable $x_1$ in $R_2$ and $R_3$ by adding multiples of $R_1$ to $R_2$ and $R_3$ to form new $R_2$ and $R_3$ .	Equivalent system of equations	Variables
	$R_1$	$x_1, x_2, x_3$
	$R_2$	$x_2, x_3$
	$R_3$	$x_2, x_3$
Eliminate variable $x_2$ in $R_3$ by adding a multiple of $R_2$ to $R_3$ to form a new $R_3$ .	Equivalent system of equations	Variables
	$R_1$	$x_1, x_2, x_3$
	$R_2$	$x_2, x_3$
	$R_3$	$x_3$

We apply this method to our given system of equations. In order to eliminate the variable  $x_1$  from  $R_2$ , we have to multiply  $R_1$  by  $-3$  and add it to  $R_2$ ; symbolically we have

$$R_2 \longmapsto R_2 + (-3R_1).$$



We then have

$$(R_1) \quad 2x_1 + x_2 - x_3 = 3$$

$$(R_2) \quad -4x_2 - 6x_3 = -2$$

$$(R_3) \quad 4x_1 + 3x_2 + x_3 = 5$$

To complete the first stage we eliminate  $x_1$  from  $R_3$ :

$$R_3 \longmapsto R_3 + (-2R_1),$$

to give

$$(R_1) \quad 2x_1 + x_2 - x_3 = 3$$

$$(R_2) \quad -4x_2 - 6x_3 = -2$$

$$(R_3) \quad x_2 + 3x_3 = -1$$

We now go to stage 2 and eliminate  $x_2$  from  $R_3$ :

$$R_3 \longmapsto R_3 + \frac{1}{4}R_2.$$

We obtain

$$(R_1) \quad 2x_1 + x_2 - x_3 = 3$$

$$(R_2) \quad -4x_2 - 6x_3 = -2$$

$$(R_3) \quad \frac{3}{2}x_3 = -\frac{3}{2}$$

This system, which is equivalent to the original system, is now in a form which can be easily solved by **back-substitution**. This means that we obtain the value of  $x_3$  from the last equation, and then substitute this value in the second equation to obtain  $x_2$ , and finally we substitute values of  $x_3$  and  $x_2$  in the first equation to obtain  $x_1$ .

$$\text{From } R_3: x_3 = -1$$

$$\text{From } R_2: 4x_2 = 2 - 6x_3 \quad \text{whence } x_2 = 2$$

$$\text{From } R_1: 2x_1 = 3 - x_2 + x_3 \quad \text{whence } x_1 = 0$$

The solution set is

$$\{(0, 2, -1)\}.$$

This is the solution set of the given system. You will note that it consists of one element only. Later we shall see just why there are no other elements in the solution set.



The following points should be noted:

- (i) The elimination method is systematic. We take no notice of any quick tricks based on the particular numbers in the equations. This is because we want a method which we can discuss theoretically and implement mechanically (on a calculating device).
- (ii) The method is deceptively simple. The deception lies in the fact that we have chosen a relatively small system and done our arithmetic exactly. In general, and especially when using a calculating machine, rounding errors will be involved which can, in unfavourable cases, cause serious trouble.
- (iii) Slight variations in the method may be necessary. For instance, there may be no  $x_1$  in the first equation. Such variations are dealt with easily in hand calculations, but require care in automatic computing.
- (iv) Gauss elimination is an *elimination* method, as opposed to an *iterative* method. In the former, we proceed step by step towards a solution, which is obtained at the *end* of the process. In the latter, we obtain an estimate of the solution at *each stage* in the process. We do not discuss iterative methods here.
- (v) As a numerical method, the method we have described has one deficiency: it has no check built into it. Of course, we can check our solution by direct substitution into the original system, but although this may tell us that we have made an error, it will not tell us where the error occurred. In fact it is quite easy to build a “running check” into the method, which checks each stage in the calculation.

We shall look at some of these points in the next chapter, when we consider numerical aspects of solving linear equations: for the present, we shall deal mainly with the theoretical aspects.

### Exercise 1

Use the Gauss elimination method to solve the following simultaneous equations:

(i)  $3x_1 - 2x_2 = 1$

$$4x_1 + x_2 = 3$$

(ii)  $x_1 - 2x_2 - x_3 = -6$

$$x_1 - 2x_2 + 2x_3 = 3$$

$$-x_1 + x_2 + x_3 = 4$$



### 7.3 Systems of Linear Equations in Matrix Form

We met matrices in Chapter 6, where they were used as convenient shorthand notation in the context of systems of equations. Now we are going to take the subject a little further. The development of notation is one of the features of mathematics. Notation usually begins as an abbreviation adopted for convenience, but may sometimes lead to significant advances as new concepts evolve around it. Matrices first appeared in about the year 1858, when Cayley introduced the notation:

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ b_m \end{pmatrix}$$

as a shorthand for the system of  $m$  simultaneous linear equations in  $n$  unknowns,  $x_1, x_2, \dots, x_n$ :

$$a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1$$

$$a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2$$

$$\cdot$$

$$\cdot$$

$$\cdot$$

$$a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m$$

You will notice that we use a double suffix notation for the coefficients. At first sight this may seem rather awesome, but it proves to be very useful. The **first suffix** specifies the **equation** to which the coefficient belongs, and the **second suffix** specifies the **variable** to which the coefficient is attached. Thus the element in the  $i$ th row and the  $j$ th column of the matrix of coefficients is  $a_{ij}$ , and  $a_{ij}$  is the coefficient of the variable  $x_j$  in the  $i$ th equation.

The significant feature of Cayley's shorthand notation is the disentanglement of the array of coefficients from the variables. This allows us to abbreviate still further by writing the system as

$$Ax = b,$$



where  $A$  is the matrix of coefficients:

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{m1} & a_{m2} & & a_{mn} \end{pmatrix},$$

and  $x$  is the (column) matrix of the  $n$  variables:

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix}.$$

The right-hand sides of the equations form the (column) matrix  $b$ :

$$b = \begin{pmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ \cdot \\ b_m \end{pmatrix}.$$

Do you feel that you have “been here before”?

By virtue of what it denotes, this notation embodies the definition of premultiplication of a column matrix by another matrix, which we denoted by  $\square$  in section 6.3. We saw that  $\square$  led to the definition of a more general operation of matrix multiplication, which we denoted by  $*$ . We have already discussed the properties of  $*$  in Chapter 6. In fact, the only new thing introduced here is the double suffix notation for the matrix elements. There is one other simplification which we have made: in accordance with general practice, we have dropped the symbol  $*$  between  $A$  and  $x$ .

Now that we have written our system of linear equations in matrix form, the solution set is a set of  $n$ -element *column matrices*, such that  $x$  belongs to the solution set if and only if

$$Ax = b.$$



We know that the solution set may be empty, or consist of one element only, or consist of more than one element.

In section 7.2 we solved the system of equations:

$$2x_1 + 1x_2 - 1x_3 = 3$$

$$6x_1 - 1x_2 - 9x_3 = 7$$

$$4x_1 + 3x_2 + 1x_3 = 5,$$

and found the solution set to be  $\{(0, 2, -1)\}$ , the set having the one element only. In matrix notation we would write the system of equations as

$$Ax = b$$

where

$$A = \begin{pmatrix} 2 & 1 & -1 \\ 6 & -1 & -9 \\ 4 & 3 & 1 \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \text{ and } b = \begin{pmatrix} 3 \\ 7 \\ 5 \end{pmatrix}.$$

The solution set consists of one element  $x$ , namely

$$x = \begin{pmatrix} 0 \\ 2 \\ -1 \end{pmatrix}.$$

In section 6.1 we showed that every system of  $m$  linear equations in  $n$  variables defines a *morphism* from the vector space  $R^n$  to the vector space  $R^m$ , and conversely, any morphism from  $R^n$  to  $R^m$  can be represented by such a system of linear equations. We saw that if we know the image of a basis of  $R^n$ , then we can find the image of *any* element of  $R^n$ , expressed as a linear combination of base vectors of  $R^m$ . In other words, we can *associate* the matrix  $A$  in the equation

$$Ax = b,$$

$A$  being a matrix of order  $m \times n$ , with a morphism

$$T: \underline{x} \longmapsto A\underline{x} \quad (\underline{x} \in R^n),$$

which maps  $R^n$  to  $R^m$ . In this text we assume that we have chosen bases for  $R^n$  and  $R^m$ , and we use these bases throughout, so that we *identify* the  $m \times n$  matrix  $A$  with the morphism from  $R^n$  to  $R^m$ . (Notice that we use  $\underline{x}$  for an element of  $R^n$  to distinguish it from the particular  $n$ -element column matrix  $x$ ; but we could just as well turn the set of all  $n$ -element



column matrices into a vector space and regard  $T$  (or  $A$ ) as mapping this space into the space of all  $m$ -element column matrices.)

So by introducing an appropriate notation for systems of linear equations, we have a bird's eye view of such systems. Instead of examining in detail *each* equation of the system, we examine the *whole system* and treat it as just one member of all possible such systems. It does not follow that this will necessarily result in any spectacular discoveries, but it may give us a better insight and understanding of these systems.

There are further advantages in using the matrix notation. For example, we can consider the matrix  $A$  to be made up of matrices of smaller order than  $A$  itself; such matrices are called *sub-matrices* of  $A$ . All these sub-matrices of  $A$  fit together to make up the matrix  $A$ . As particular examples of this, we can think of the  $n$  columns of elements of the  $(m \times n)$  matrix  $A$  as  $n$  column vectors or as  $n$  submatrices of order  $m \times 1$ . Similarly, we can think of the matrix  $A$  as one column of  $m$  row vectors, each row vector having  $n$  components (i.e. the components of the corresponding row of the matrix  $A$ ).

### Example 1

We can write the matrix

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & -2 & 7 \end{pmatrix}$$

as

$$A = (b_1, b_2, b_3),$$

where  $b_1, b_2, b_3$  are the matrices

$$b_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad b_2 = \begin{pmatrix} 2 \\ -2 \end{pmatrix} \text{ and } b_3 = \begin{pmatrix} 3 \\ 7 \end{pmatrix}.$$

Also we can write  $A$  as

$$A = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix},$$

where  $c_1, c_2$  are the matrices

$$c_1 = (1 \quad 2 \quad 3) \text{ and } c_2 = (2 \quad -2 \quad 7).$$

We call  $c_1$  and  $c_2$  **row matrices** to distinguish them from the **column matrices**  $b_1, b_2$  and  $b_3$ , which are the columns of the matrix  $A$ . Both row



and column matrices can be regarded as  $n$ - and  $m$ -tuples of numbers, and we can regard them as elements of vector spaces isomorphic to  $R^n$  and  $R^m$  respectively. Hence the names row and column *vectors*.

## 7.4 The Nature of the Solution II

In the last section we referred back to the work on vector spaces which we covered in Chapter 5. In this and the following section, we shall again make use of some of the results mentioned there.

Our problem is to solve a system of linear equations, which in matrix form can be written

$$A\underline{x} = \underline{b},$$

where  $A$  is a matrix of order  $m \times n$ .  $A$  defines a *morphism* from  $R^n$  to  $R^m$ , also denoted by  $A$ :

$$A:\underline{x} \longmapsto A\underline{x} \quad (\underline{x} \in R^n).$$

(Notice that we now underline  $\underline{x}$  and  $\underline{b}$ , to emphasize that we are considering them as *vectors*, i.e. elements of a vector space.)

We know from section 5.6 that we can obtain the required solution set in two parts.

- (i) We need just *one element* of the actual solution set, i.e. one vector  $\underline{x}$  which satisfies the equation

$$A\underline{x} = \underline{b}. \quad \text{Equation (1)}$$

- (ii) We need the *complete* solution set of the equation

$$A\underline{x} = \underline{0}, \quad \text{Equation (2)}$$

where  $\underline{0}$  is the zero vector (i.e. the column matrix all of whose elements are zero) in  $R^m$ . This solution set is the *kernel* of the morphism.

The actual solution set of the original system is then obtained by adding the solution in (i) to each solution in (ii).

We shall verify this result again in our particular circumstances to remind you of the argument.

Suppose that  $\underline{x}_p$  is a solution of Equation (1), and  $\underline{x}_k$  is a member of the kernel,  $K$  (i.e. a solution of Equation (2)); then

$$\begin{aligned} A(\underline{x}_p + \underline{x}_k) &= A\underline{x}_p + A\underline{x}_k && (A \text{ is a morphism}) \\ &= \underline{b} + \underline{0} && (\text{hypothesis}) \\ &= \underline{b} && (\text{definition of } \underline{0}) \end{aligned}$$



This shows that  $(\underline{x}_p + \underline{x}_k)$  is an element of the solution set of Equation (1).

Also, every element of the solution set of Equation (1) can be written in the form  $\underline{x}_p + \underline{x}_k$ . For if we let  $\underline{x}_1$  be any element of that solution set, then

$$\begin{aligned} A(\underline{x}_1 - \underline{x}_p) &= A\underline{x}_1 - A\underline{x}_p \\ &= \underline{b} - \underline{b} \\ &= \underline{0} \end{aligned}$$

It follows that the vector  $(\underline{x}_1 - \underline{x}_p)$  is a solution of Equation (2), i.e. an element of  $K$ , so that we can write

$$\underline{x}_1 - \underline{x}_p = \underline{x}_k,$$

where  $\underline{x}_k \in K$ , and so

$$\underline{x}_1 = \underline{x}_p + \underline{x}_k.$$

So it follows that there is a one-one correspondence between the solution set and the kernel. Thus, the solution set is

$$\{\underline{x}_i : \underline{x}_i = \underline{x}_p + \underline{x}_k, \underline{x}_k \in K\}.$$

$\underline{x}_p$  is called a **particular solution** of  $A\underline{x} = \underline{b}$ .

### Example 1

In Example 5.6.3 we considered the system of equations

$$2x + 3y + (-1)z = 1$$

$$1x + 1y + (-1)z = 2$$

In matrix form it becomes

$$\begin{pmatrix} 2 & 3 & -1 \\ 1 & 1 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

We found that a particular solution is

$$\underline{x}_p = \begin{pmatrix} 5 \\ -3 \\ 0 \end{pmatrix},$$



and the kernel is the set

$$K = \left\{ \begin{pmatrix} 2r \\ -r \\ r \end{pmatrix} : r \in R \right\}.$$

The solution set of the system in matrix or column vector form is therefore:

$$\left\{ \begin{pmatrix} 5 + 2r \\ -3 - r \\ 0 + r \end{pmatrix} : r \in R \right\}.$$

### Exercise 1

Find the matrix form of the solution set of the system

$$x + y + z = 4$$

$$2x + 2y - z = 5$$

### Existence and Uniqueness Problems

In section 7.1 we noticed that some systems have an empty solution set, whereas others have a non-empty solution set. One of the important problems in the theory of linear equations is to determine conditions under which a system of equations has a non-empty solution set: this problem is called the **existence problem**.

Once it has been decided whether or not a solution exists, it is useful to know the conditions under which the solution set contains just *one* element: this problem is known as the **uniqueness problem**.

We have seen that the solution set of the equation

$$A\underline{x} = \underline{b}$$

is

$$\{\underline{x} : \underline{x} = \underline{x}_p + \text{any element of the kernel}\}.$$

Thus, if we can find a particular solution  $\underline{x}_p$  and determine the nature of the kernel of the mapping, the uniqueness problem is solved.

If the kernel consists of one element only (which must then be the zero element), then the solution set consists of one element only,  $\underline{x}_p$ , and so we



have established uniqueness. If, on the other hand, the kernel consists of more than one element, then the system has more than one solution.

As far as existence is concerned, we are assured of a solution if  $\underline{b}$  is in the image set of the mapping defined by  $A$ , for in that case there must be some vector  $\underline{x}$  which maps to  $\underline{b}$ .

Thus the existence problem is essentially the problem of finding a test by which we can look at  $A$  and  $\underline{b}$  and determine whether  $\underline{b}$  is in the image set. If the solution set is non-empty, the uniqueness problem will be solved if we can find a way of determining whether the kernel of the mapping defined by  $A$  contains more than one element. In the following sections we shall consider these problems in more detail.

## 7.5 The Existence Problem

We consider a system of  $m$  simultaneous equations in  $n$  unknowns,  $x_1, x_2, \dots, x_n$ , written in matrix form as

$$A\underline{x} = \underline{b},$$

where

$$\underline{x} = \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix}, \quad \underline{b} = \begin{pmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ \cdot \\ b_m \end{pmatrix},$$

and  $A$  is a matrix of order  $m \times n$ :

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{m1} & a_{m2} & & a_{mn} \end{pmatrix}$$



In section 7.3 we remarked that we can consider a matrix to be made up of vectors. For example, we could write

$$\underline{a}_1 = \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{pmatrix}, \quad \underline{a}_2 = \begin{pmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{pmatrix}, \dots, \underline{a}_n = \begin{pmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{pmatrix}$$

If we choose the following basis for  $R^n$ :

$$\underline{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \underline{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \dots, \underline{e}_n = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix},$$

then

$$\underline{x} = x_1 \underline{e}_1 + x_2 \underline{e}_2 + \dots + x_n \underline{e}_n$$

and

$$\begin{aligned} A\underline{x} &= x_1 A\underline{e}_1 + x_2 A\underline{e}_2 + \dots + x_n A\underline{e}_n \\ &= x_1 \underline{a}_1 + x_2 \underline{a}_2 + \dots + x_n \underline{a}_n. \end{aligned}$$

### Example 1

If we write the system of equations

$$2x_1 + 3x_2 = 1$$

$$5x_1 + 1x_2 = 3$$

in matrix form:

$$\begin{pmatrix} 2 & 3 \\ 5 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \end{pmatrix},$$

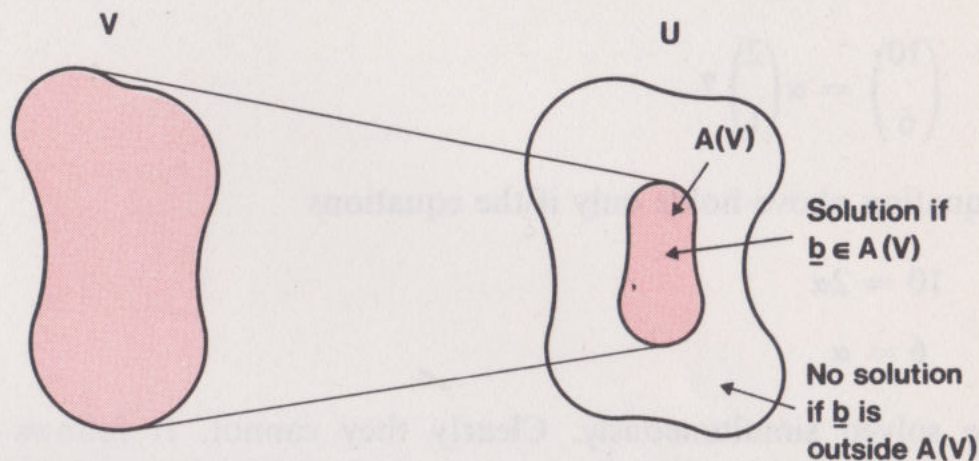
then we can write this as

$$x_1 \begin{pmatrix} 2 \\ 5 \end{pmatrix} + x_2 \begin{pmatrix} 3 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \end{pmatrix}.$$

This illustrates the general formula given above.



Now let us return to our problem of the existence of a non-empty solution set.  $A\underline{x}$  is a vector in the image space  $A(R^n)$ . Since  $A\underline{x} = x_1\underline{a}_1 + x_2\underline{a}_2 + \cdots + x_n\underline{a}_n$ , this means that any vector in the image space is a linear combination of the vectors  $\underline{a}_1, \underline{a}_2, \dots, \underline{a}_n$ . For a solution to exist,  $\underline{b}$  must be in  $A(R^n)$ , and so  $\underline{b}$  must be a linear combination of  $\underline{a}_1, \underline{a}_2, \dots, \underline{a}_n$ .



So we have what appears to be a simple test which we can apply to find whether a system has a solution. Unfortunately the test is not always simple to apply in practice.

### The Test

If  $\underline{b}$  is a linear combination of the vectors  $\underline{a}_1, \underline{a}_2, \dots, \underline{a}_n$ , then the system of linear equations represented by  $A\underline{x} = \underline{b}$  has a solution. Otherwise the system has no solution (that is, the solution set is empty).

### Example 2

Consider the system of equations represented by

$$\begin{pmatrix} 2 & 4 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 10 \\ 6 \end{pmatrix}.$$

We must decide whether the vector  $\underline{b} = \begin{pmatrix} 10 \\ 6 \end{pmatrix}$  is a linear combination of the 2 vectors

$$\underline{a}_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \quad \text{and} \quad \underline{a}_2 = \begin{pmatrix} 4 \\ 2 \end{pmatrix}.$$

In fact, we notice that

$$\underline{a}_2 = 2\underline{a}_1, \text{ since } \begin{pmatrix} 4 \\ 2 \end{pmatrix} = 2 \times \begin{pmatrix} 2 \\ 1 \end{pmatrix},$$



so that effectively our problem reduces to deciding whether  $\underline{b}$  is a (scalar) multiple of  $\underline{a}_1$ . In other words, is there a number  $\alpha$  such that

$$\underline{b} = \alpha \underline{a}_1,$$

i.e.

$$\begin{pmatrix} 10 \\ 6 \end{pmatrix} = \alpha \begin{pmatrix} 2 \\ 1 \end{pmatrix} ?$$

The equation above holds only if the equations

$$10 = 2\alpha$$

$$6 = \alpha$$

can be solved simultaneously. Clearly they cannot. It follows that  $\underline{b}$  is *not* a linear combination of  $\underline{a}_1$  and  $\underline{a}_2$ , so that there is *no solution* to the original system of equations.

### Exercise 1

Using the test given in the text, determine which of the following systems of equations have at least one solution.

(i)  $x_1 + 2x_2 = 5$

(ii)  $x_1 + 2x_2 = 3$

$x_1 + x_2 = 3$

$3x_1 + 6x_2 = 9$

(iii)  $\frac{1}{2}x_1 + \frac{1}{3}x_2 = 2$

(iv)  $0.2x_1 + 0.3x_2 + 0.1x_3 = 1.1$

$\frac{3}{2}x_1 + x_2 = 8$

$0.6x_1 + 0.9x_2 + 0.3x_3 = 2.2.$

In the examples considered so far, the test to determine whether the system  $A\underline{x} = \underline{b}$  has a solution was easy to apply. In fact, for a “small” system we do not need a test; we can establish whether or not a solution exists by trying to solve the equations directly. But for “meatier” systems a test is required. However, in such a case, to find whether  $\underline{b}$  is a linear combination of the columns of  $A$ , we would need to solve a system of simultaneous equations, which could be as complicated as the original system! Clearly, we must modify the test to make it easier to apply. We do this by introducing the concept of the *rank* of a matrix.

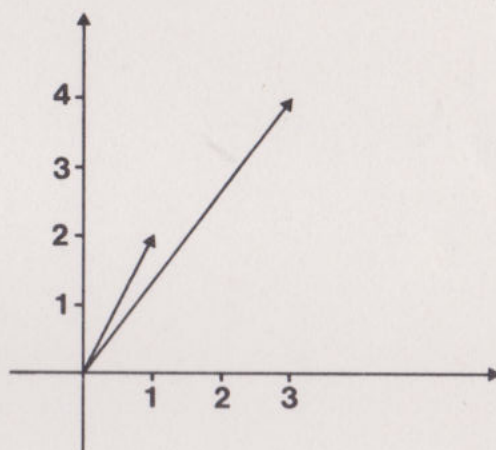
Given the matrix  $A = (\underline{a}_1 \ \underline{a}_2 \ \dots \ \underline{a}_n)$ , the **rank of  $A$**  is the maximum number of linearly independent vectors from the set  $\{\underline{a}_1, \underline{a}_2, \dots, \underline{a}_n\}$ ; it is denoted by  $r(A)$ .



**Example 3**

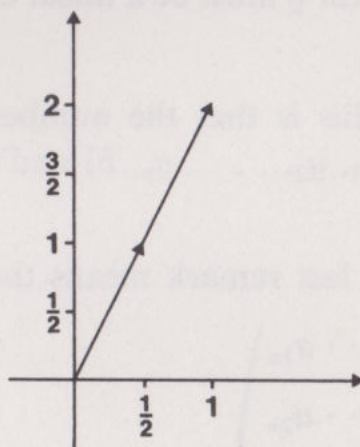
The matrix  $\begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}$  has rank 2, since the vectors  $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$  and  $\begin{pmatrix} 3 \\ 4 \end{pmatrix}$  are linearly independent:

$$\alpha_1 \begin{pmatrix} 1 \\ 2 \end{pmatrix} + \alpha_2 \begin{pmatrix} 3 \\ 4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ implies } \alpha_1 = \alpha_2 = 0.$$



The matrix  $\begin{pmatrix} 1 & \frac{1}{2} \\ 2 & 1 \end{pmatrix}$  has rank 1, since the vectors  $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$  and  $\begin{pmatrix} \frac{1}{2} \\ 1 \end{pmatrix}$  are linearly dependent:

$$\frac{1}{2} \begin{pmatrix} 1 \\ 2 \end{pmatrix} + (-1) \begin{pmatrix} \frac{1}{2} \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$



We have defined rank in terms of the number of linearly independent column vectors, because our discussion has been in these terms. We could equally well consider the matrix to be made up of *row* vectors, and define



rank in terms of the number of linearly independent row vectors. (In a sense, this would be quite natural, since intuitively it connects with the number of “independent” equations in the system.) There is a theorem (which we shall not prove) which states that these two possible definitions of the rank of a matrix are equivalent, i.e. give the same value for the rank.

### Exercise 2

Find the rank of each of the following matrices.

$$(i) \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}$$

$$(ii) \begin{pmatrix} 2 & 1 & -1 \\ 4 & 1 & 1 \\ 6 & 1 & 3 \end{pmatrix}$$

$$(iii) \begin{pmatrix} 6 & 3 & 9 \\ 2 & 1 & 3 \\ 4 & 2 & 6 \end{pmatrix}$$

We now reconsider our test for the existence of a non-empty solution set in terms of the rank concept.

We have seen that for the system

$$A\underline{x} = \underline{b}, \text{ where } A = (\underline{a}_1 \dots \underline{a}_n)$$

to have a solution, the vector  $\underline{b}$  must be a linear combination of the vectors

$$\underline{a}_1, \underline{a}_2, \dots, \underline{a}_n.$$

Another way of saying this is that the number of linearly independent vectors in the two sets  $\{\underline{a}_1, \underline{a}_2, \dots, \underline{a}_n, \underline{b}\}$  and  $\{\underline{a}_1, \underline{a}_2, \dots, \underline{a}_n\}$  must be the same.

In terms of matrices, the last remark means that the rank of the matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$



and the rank of the **augmented matrix** (the matrix  $A$  with the extra column  $\underline{b}$ ):

$$(A \ \underline{b}) = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ a_{m1} & a_{m2} & \cdots & a_{mn} & b_m \end{pmatrix}$$

must be the same.

We can now phrase the test for the existence of a non-empty solution set in the following form:

The system of linear equations represented by

$$A\underline{x} = \underline{b}$$

has a non-empty solution set if the matrix  $A$  and the augmented matrix  $(A \ \underline{b})$  have the same rank.

Unfortunately, with our present techniques, to find whether  $r(A)$  is the same as  $r(A \ \underline{b})$  may still necessitate solving a system of equations which is as large as the original system we wish to solve. We shall discuss another possible method for determining the rank of a matrix (which is not dependent on solving simultaneous equations) in the next chapter.

## 7.6 The Uniqueness Problem

Having discussed some theoretical aspects of the existence problem, we now turn our attention to the uniqueness problem.

In general, corresponding to the equation

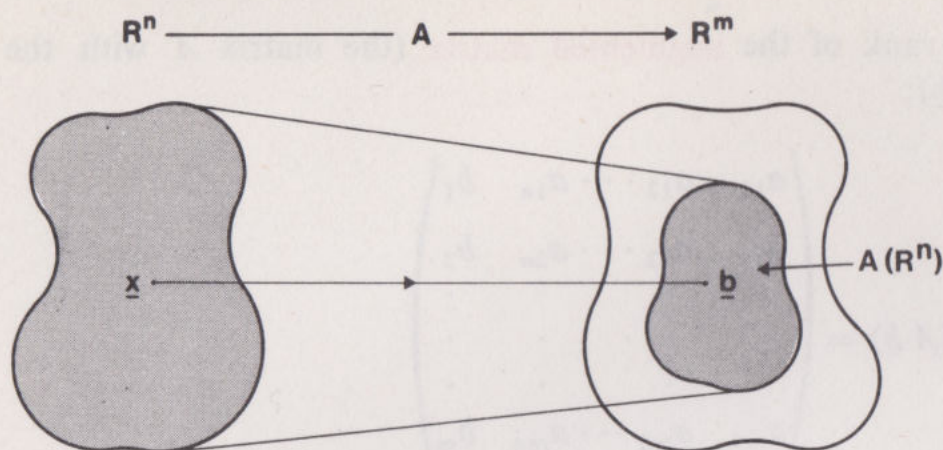
$$A\underline{x} = \underline{b},$$

where  $A$  has order  $m \times n$ , we have a mapping

$$A: \underline{x} \longmapsto A\underline{x}$$

of  $R^n$  to  $R^m$ . We know that, for the equation  $A\underline{x} = \underline{b}$  to have a solution,  $\underline{b}$  must be a linear combination of the vectors  $\underline{a}_1, \underline{a}_2, \dots, \underline{a}_n$ . So we can determine the image set  $A(R^n) \subseteq R^m$ , as the set spanned by  $\underline{a}_1, \underline{a}_2, \dots, \underline{a}_n$ .





The dimension of this image set is, therefore, the maximum number of linearly independent vectors among the  $\underline{a}$ 's. But this is what we have defined to be the rank of  $A$ , so we have

$$r(A) = \text{dimension of } A(\mathbb{R}^n).$$

We can get some interesting results by recalling the dimension theorem of section 5.6. This states, in terms of our context, that

dimension of  $A(\mathbb{R}^n) = \text{dimension of } \mathbb{R}^n - \text{dimension of kernel,}$   
i.e.

$$r(A) = n - \text{dimension of kernel.}$$

Now let us suppose that the solution of  $A\underline{x} = \underline{b}$  is unique. We know in general that the set of all solutions is given by

$$\{\underline{x} : \underline{x} = \underline{x}_p + \underline{x}_k, \underline{x}_k \in K\},$$

where  $\underline{x}_p$  is a particular solution and  $K$  is the kernel of the mapping  $A$ . But if, as we suppose, the solution is unique, then there is only one solution, so the kernel contains just the one element, the zero vector.<sup>†</sup> This means that the dimension of the kernel is zero, i.e. our result above now becomes

$$r(A) = n.$$

Thus a *necessary* condition for the solution of  $A\underline{x} = \underline{b}$  to be unique is that the rank of  $A = n$ , the number of variables in the original equations. This condition is obviously also *sufficient*, i.e. it guarantees uniqueness. Because if  $r(A) = n$ , then the dimension theorem tells us that the dimension of the kernel is zero, and therefore the kernel is the zero vector space. It follows that, if  $A\underline{x} = \underline{b}$  has a solution, then it is unique. So we have:

If a system of  $m$  linear equations in  $n$  unknowns represented by

$$A\underline{x} = \underline{b}$$

<sup>†</sup> Since  $\alpha \underline{0} = \underline{0}$ , where  $\underline{0}$  is the zero vector in a vector space, and  $\alpha \neq 0$ ,  $\{\underline{0}\}$  is linearly dependent. The dimension of  $\{\underline{0}\}$  is defined to be zero. It is the only vector space of dimension zero.



has a solution, then the solution is unique if and only if

$$r(A) = n.$$

We can improve on this result if we simplify our case. For, if we assume that  $n = m$ , i.e. the number of equations is equal to the number of variables, then we can include existence with uniqueness. That is,

The system of  $n$  equations in  $n$  unknowns represented by

$$A\underline{x} = \underline{b}$$

has a unique solution if and only if

$$r(A) = n.$$

We have already shown that, if a solution exists, it is unique, and to prove existence is not difficult. Since  $n = m$ , the image set  $A(R^n)$  is now a subset of  $R^n$ . Since the column vectors  $\underline{a}_1, \underline{a}_2, \dots, \underline{a}_n$  are linearly independent, it can be shown that they form a basis for  $R^n$ . (We stated this result in section 5.3, but we shall not prove it in this volume.) This means that any  $\underline{b} \in R^n$  can be expressed as a linear combination of  $\underline{a}_1, \underline{a}_2, \dots, \underline{a}_n$ . It follows that  $A\underline{x} = \underline{b}$  has a solution.

In section 7.5 we showed, in general and not only when  $n = m$ , that a *sufficient* condition for the existence of a solution of

$$A\underline{x} = \underline{b}$$

is that

$$r(A) = r(A \ \underline{b}).$$

We have now shown that when  $n = m$ , a *necessary* and *sufficient* condition for the existence of a unique solution is that

$$r(A) = n.$$

Let us consider the case where  $m = n$ , so that we can compare the above two results.

Suppose

$$r(A) = n.$$

Then

$$r(A \ \underline{b}) \geq r(A) = n.$$

Since the columns of  $(A \ \underline{b})$  are elements of a vector space of dimension  $n$ , it follows that at most  $n$  of them are linearly independent, i.e.

$$r(A \ \underline{b}) \leq n.$$



Hence

$$r(A \ \underline{b}) = n.$$

That is,

$$r(A) = n \text{ implies } r(A) = r(A \ \underline{b}).$$

On the other hand,

$$r(A) = r(A \ \underline{b}) \text{ does not imply } r(A) = n.$$

### Example 1

The system

$$x - y = 2$$

$$x + y = 2$$

has the *unique* solution  $x = 2, y = 0$ .

Here

$$A = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}, \quad (A \ \underline{b}) = \begin{pmatrix} 1 & -1 & 2 \\ 1 & 1 & 2 \end{pmatrix}$$

and

$$r(A) = r(A \ \underline{b}) = 2.$$

The system

$$2x + 2y = 2$$

$$x + y = 1$$

has *many* solutions of the form  $x = \alpha, y = 1 - \alpha$ .

In this case,

$$A = \begin{pmatrix} 2 & 2 \\ 1 & 1 \end{pmatrix}, \quad (A \ \underline{b}) = \begin{pmatrix} 2 & 2 & 2 \\ 1 & 1 & 1 \end{pmatrix}$$

and

$$r(A) = r(A \ \underline{b}) = 1 < 2.$$

We have now completed, as far as we intend to go, our theoretical studies of existence and uniqueness. But before we go on to other things, we introduce a few terms and notation which are standard in the literature on linear algebra.



When the number of equations is equal to the number of variables, the matrix of coefficients is said to be **square**. A matrix with  $m$  rows and  $n$  columns is often said to be of **order**  $m \times n$ . The square matrix  $A$  of order  $n \times n$  (or sometimes “of order  $n$ ”) is said to be **non-singular** if  $r(A) = n$ . Otherwise  $A$  is called **singular**.

In general, the mapping

$$A: \underline{x} \longmapsto A\underline{x}$$

is a homomorphism, but if the matrix is non-singular, the mapping is an isomorphism of  $R^n$  to  $R^n$ . In this case, the mapping has an inverse which is also an isomorphism. We denote the matrix of the inverse mapping, and also the inverse mapping itself, by  $A^{-1}$ . In terms of mappings, we have

$$A^{-1} \circ A: \underline{x} \longmapsto \underline{x}$$

and

$$A \circ A^{-1}: \underline{x} \longmapsto \underline{x}.$$

If  $I_{n,n}$  denotes the identity matrix of order  $n$  (see section 6.4), then in terms of matrices we have

$$A^{-1}A = AA^{-1} = I_{n,n}.$$

$A^{-1}$  is called the **inverse matrix** of  $A$ .

If

$$A\underline{x} = \underline{b}$$

has a unique solution,  $\underline{x}_p$ , then we can write

$$\underline{x}_p = A^{-1} \underline{b}.$$

In the next chapter we shall see how to calculate  $A^{-1}$ , and then this formula can prove useful; for instance, we may want the solutions of several sets of equations with the *same*  $A$ , but various  $\underline{b}$ .

## 7.7 Summary

In section 7.3 we introduced the matrix form of a system of simultaneous linear equations:

$$A\underline{x} = \underline{b}.$$

In section 7.4 we gave a general discussion of the nature of the solution. In particular, we mentioned the *existence problem*:

Does a solution *exist*?



and the *uniqueness problem*:

Is there a *unique* solution?

In section 7.5 we discussed the existence problem in detail. We defined the *rank* of a matrix:

rank of  $A$ ,  $r(A) = (\text{maximum number of linearly independent columns of } A)$

and the *augmented matrix*:

$$(A \ \underline{b}) = \left( A \mid \underline{b} \right).$$

We gave the following theorem:

$$r(A) = r(A \ \underline{b})$$

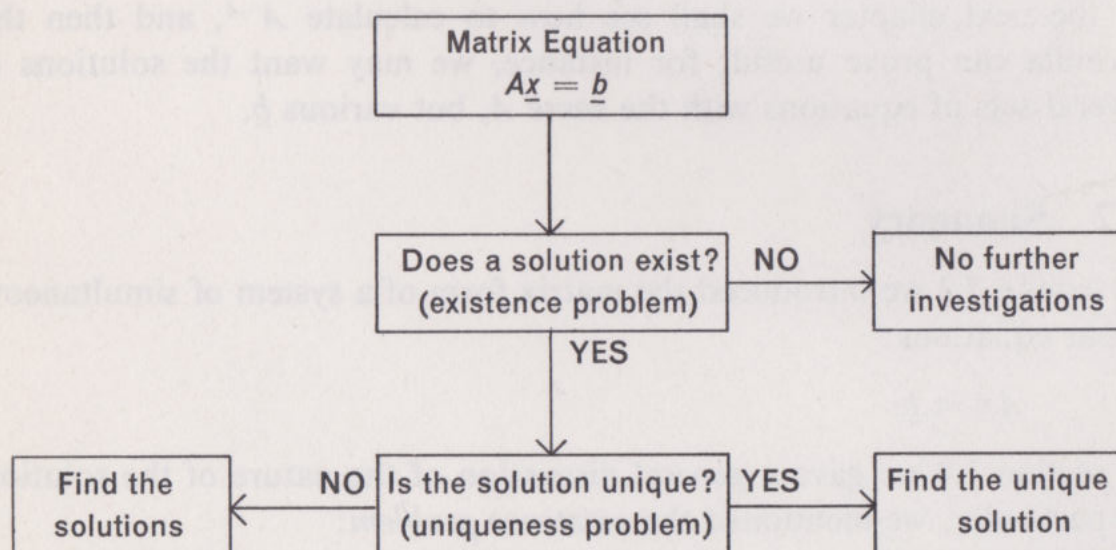
implies  $A\underline{x} = \underline{b}$  has a non-empty solution set.

In section 7.6 we discussed the connection between the uniqueness problem and the rank of a matrix; we produced the following theorem:

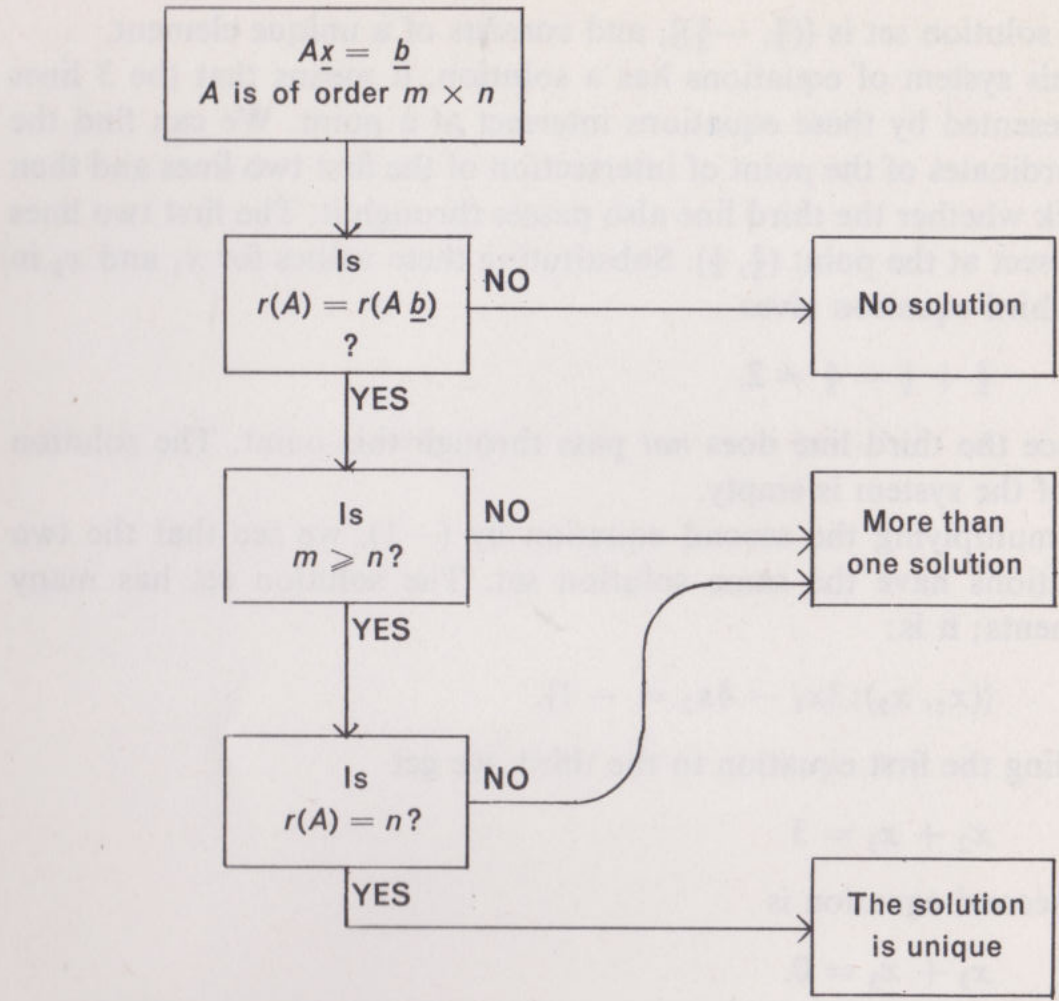
If  $A\underline{x} = \underline{b}$  is a system of linear equations in  $n$  unknowns, for which a solution exists, then

the solution is unique if and only if  $r(A) = n$ .

A summary of the discussion on the existence and uniqueness problems is given below in diagrammatic form.







Since this is basically a theoretical chapter, we have set no additional exercises. In the next chapter we shall apply the theory and set a number of numerical exercises.

7.8 Answers to Exercises

Section 7.1

Exercise 1

	A	B	C
(i)		✓	
(ii)	✓		
(iii)			✓
(iv)	✓		



- (i) The solution set is  $\{(\frac{7}{3}, -\frac{5}{3})\}$ , and consists of a unique element.
- (ii) If this system of equations has a solution, it means that the 3 lines represented by these equations intersect at a point. We can find the co-ordinates of the point of intersection of the first two lines and then check whether the third line also passes through it. The first two lines intersect at the point  $(\frac{3}{5}, \frac{1}{5})$ . Substituting these values for  $x_1$  and  $x_2$  in the third equation gives

$$\frac{3}{5} + \frac{1}{5} = \frac{4}{5} \neq 2.$$

Hence the third line does *not* pass through this point. The solution set of the system is empty.

- (iii) On multiplying the second equation by  $(-1)$ , we see that the two equations have the same solution set. The solution set has many elements; it is:

$$\{(x_1, x_2): 3x_1 - 4x_2 = -1\}.$$

- (iv) Adding the first equation to the third, we get

$$x_2 + x_3 = 3$$

and the second equation is

$$x_3 + x_3 = 0.$$

These equations cannot be satisfied simultaneously; there is no solution.

## Section 7.2

### Exercise 1

- (i) The solution set is  $\{(\frac{7}{11}, \frac{5}{11})\}$ .
- (ii) Eliminating  $x_1$  from  $R_2$  and  $R_3$ , we obtain

$$x_1 - 2x_2 - x_3 = -6$$

$$3x_3 = 9$$

$$-x_2 = -2$$

It is unnecessary to carry on further, since the solutions can be found by back substitution:

$$x_2 = 2$$

$$x_3 = 3$$

$$x_1 = -6 + x_3 + 2x_2 = 1$$

Hence the solution set is  $\{(1, 2, 3)\}$ .



## Section 7.4

## Exercise 1

If we try to find a particular solution by putting  $z = 0$ , we obtain

$$x + y = 4$$

$$2x + 2y = 5.$$

We see that there is *no* particular solution of the original system of the form

$$\begin{pmatrix} \alpha \\ \beta \\ 0 \end{pmatrix}.$$

We therefore try putting another variable,  $y$  say, equal to zero; we obtain the system

$$x + z = 4$$

$$2x - z = 5,$$

which has the unique solution  $x = 3$  and  $z = 1$ . So

$$\begin{pmatrix} 3 \\ 0 \\ 1 \end{pmatrix}$$

is a particular solution of the original system. We now wish to find the kernel, i.e. to solve the system

$$x + y + z = 0$$

$$2x + 2y - z = 0.$$

The solution set of this system is

$$\left\{ \begin{pmatrix} r \\ -r \\ 0 \end{pmatrix} : r \in R \right\}.$$

Hence the required solution set is

$$\left\{ \begin{pmatrix} 3 + r \\ -r \\ 1 \end{pmatrix} : r \in R \right\}.$$



## Section 7.5

## Exercise 1

(i) The system  $A\underline{x} = \underline{b}$  is

$$\begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 5 \\ 3 \end{pmatrix}.$$

By inspection,

$$\underline{b} = \begin{pmatrix} 5 \\ 3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} + 2 \begin{pmatrix} 2 \\ 1 \end{pmatrix},$$

so  $\underline{b}$  is a linear combination of the column vectors of the matrix of coefficients  $A$ . It follows that the system has at least one solution.

(ii) The system  $A\underline{x} = \underline{b}$  is

$$\begin{pmatrix} 1 & 2 \\ 3 & 6 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 3 \\ 9 \end{pmatrix}.$$

In this case,

$$\underline{b} = \begin{pmatrix} 3 \\ 9 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \end{pmatrix} + \begin{pmatrix} 2 \\ 6 \end{pmatrix}.$$

It follows that the system has at least one solution.

(iii) The system  $A\underline{x} = \underline{b}$  is

$$\begin{pmatrix} \frac{1}{2} & \frac{1}{3} \\ \frac{3}{2} & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 8 \end{pmatrix}.$$

The two column vectors of  $A$  are linearly dependent:

$$\begin{pmatrix} \frac{1}{2} \\ \frac{3}{2} \end{pmatrix} = \frac{3}{2} \times \begin{pmatrix} \frac{1}{3} \\ 1 \end{pmatrix}$$

So the system will only have a solution if the vector  $\underline{b}$  is a multiple

of (say) the vector  $\begin{pmatrix} \frac{1}{3} \\ 1 \end{pmatrix}$ , i.e. if

$$\begin{pmatrix} 2 \\ 8 \end{pmatrix} = \alpha \begin{pmatrix} \frac{1}{3} \\ 1 \end{pmatrix}, \text{ where } \alpha \in R.$$

There is no  $\alpha$  such that the above equation holds. The system has no solution.



(iv) The system  $A\underline{x} = \underline{b}$  is

$$\begin{pmatrix} 0.2 & 0.3 & 0.1 \\ 0.6 & 0.9 & 0.3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1.1 \\ 2.2 \end{pmatrix}.$$

The three column vectors of  $A$ ,  $\underline{a}_1$ ,  $\underline{a}_2$ ,  $\underline{a}_3$ , are linearly dependent; in fact,

$$\underline{a}_1 = 2 \times \underline{a}_3 \quad \text{and} \quad \underline{a}_2 = 3 \times \underline{a}_3.$$

But the vector  $\underline{b}$  is not a multiple of  $\underline{a}_3$ . It follows that the system has no solution.

### Exercise 2

$$(i) \quad \alpha_1 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + \alpha_3 \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\text{whence} \quad \alpha_1 + \alpha_2 = 0$$

$$\alpha_3 = 0$$

$$\alpha_1 + \alpha_3 = 0$$

it follows that  $\alpha_1 = \alpha_2 = \alpha_3 = 0$ .

This shows that the vectors  $\underline{a}_1$ ,  $\underline{a}_2$  and  $\underline{a}_3$  are linearly independent, and hence  $r(A) = 3$ .

$$(ii) \quad \underline{a}_1 = \begin{pmatrix} 2 \\ 4 \\ 6 \end{pmatrix}, \underline{a}_2 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \underline{a}_3 = \begin{pmatrix} -1 \\ 1 \\ 3 \end{pmatrix}. \text{ You can verify that } 3\underline{a}_2 + \underline{a}_3 = \underline{a}_1.$$

It follows that the maximum number of linearly independent column vectors is less than 3. Looking at the expression

$$\alpha_2 \underline{a}_2 + \alpha_3 \underline{a}_3 = \underline{0},$$



we see that this is equivalent to

$$\alpha_2 - \alpha_3 = 0$$

$$\alpha_2 + \alpha_3 = 0$$

$$\alpha_2 + 3\alpha_3 = 0$$

The first two equations give  $\alpha_2 = 0$ , and it then follows that  $\alpha_3 = 0$ . Thus  $\underline{a}_2$  and  $\underline{a}_3$  are linearly independent, so  $r(A) = 2$ .

(iii)  $\underline{a}_1 = 2\underline{a}_2$ ,  $\underline{a}_3 = 3\underline{a}_2$ .

In this case the maximum number of linearly independent column vectors is 1, so that  $r(A) = 1$ .



## CHAPTER 8 NUMERICAL METHODS

### 8.0 Introduction

In section 7.2 we discussed one practical method for solving a system of equations, the Gauss elimination method. We then used the matrix notation, which is convenient for investigating the problems of the existence and uniqueness of solutions. We shall begin this chapter by looking at the solution of a system of equations using the Gauss elimination method, but in terms of matrices. This has no practical advantage; in fact, it is a disadvantage. But it does allow us to discuss certain theoretical aspects of the method which have definite practical repercussions. We shall also be looking at the problem of calculating the inverse of a matrix and its rank.

The Gauss elimination method is just one of a number of methods for solving a system of linear equations. In sections 8.4 and 8.5 we investigate various methods.

Finally, we come to the problem of accuracy. Frequently, the data we use to set up the equations are inaccurate, and so the eventual result may be in error, not only because of round-off errors which may have built up during the computation, but also from inaccuracies propagated from the very start. In certain special cases this inaccuracy makes the results almost worthless. It is this aspect of the solution of simultaneous equations, called *ill-conditioning*, which we shall discuss at the end of this chapter.

### 8.1 Elementary Matrices

In section 7.2 we defined three *elementary operations* used in the Gauss elimination method. We shall show that the equivalent operations, when the equations are written in matrix form, are multiplications of the coefficient matrix  $A$  by appropriate matrices. For simplicity we shall confine our attention to matrices of order  $3 \times 3$ , but the method we discuss is very general and can be applied to matrices of any order.

We define three **elementary row operations** on a matrix  $A$ :

- (i) interchange any two rows of the matrix;
- (ii) multiply any row of the matrix by a non-zero number;
- (iii) add a multiple of one row to another row.

We shall denote the first, second and third rows of the matrix  $A$  by  $R_1$ ,  $R_2$  and  $R_3$  respectively.



We denote an elementary row operation by  $E_i$  and abbreviate the descriptions as follows.

Interchange of  $R_1$  and  $R_3$  is written

$$E_1: R_1 \longleftrightarrow R_3.$$

Multiplication of  $R_1$  by the number  $k$  is written

$$E_2: R_1 \longmapsto kR_1.$$

Addition of a multiple,  $k$ , of  $R_3$  to  $R_1$  is written

$$E_3: R_1 \longmapsto R_1 + kR_3.$$

### Example 1

Consider

$$A = \begin{pmatrix} 1 & 0 & 3 \\ 1 & 1 & 1 \\ 1 & 0 & -2 \end{pmatrix}.$$

$E_1: R_1 \longleftrightarrow R_2$  changes  $A$  to

$$E_1(A) = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 3 \\ 1 & 0 & -2 \end{pmatrix}.$$

$E_2: R_2 \longmapsto 3R_2$  changes  $A$  to

$$E_2(A) = \begin{pmatrix} 1 & 0 & 3 \\ 3 & 3 & 3 \\ 1 & 0 & -2 \end{pmatrix},$$

and changes  $E_1(A)$  to

$$E_2(E_1(A)) = \begin{pmatrix} 1 & 1 & 1 \\ 3 & 0 & 9 \\ 1 & 0 & -2 \end{pmatrix}.$$

$E_3: R_2 \longmapsto R_2 - 5R_3$  changes  $A$  to

$$\begin{pmatrix} 1 & 0 & 3 \\ -4 & 1 & 11 \\ 1 & 0 & -2 \end{pmatrix},$$



and changes  $E_2(A)$  to

$$E_3(E_2(A)) = \begin{pmatrix} 1 & 0 & 3 \\ -2 & 3 & 13 \\ 1 & 0 & -2 \end{pmatrix}.$$

It is a remarkable fact that these operations can be performed on a matrix  $A$  by premultiplying  $A$  by particular matrices. The most direct method of demonstrating this is to find the matrices which perform the required operations. How do we find these matrices? There is a particularly simple way to do this. If we *assume* that such matrices exist, then they must perform the same operations on the identity matrix, in particular. For example, if  $E$  is *assumed* to be a matrix which interchanges the first and third rows of a  $3 \times 3$  matrix, then

$$E = EI = E \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix},$$

since we know that  $E$  interchanges the first and third rows. Now let us premultiply a general matrix by  $E$ :

$$\begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \begin{pmatrix} a_{31} & a_{32} & a_{33} \\ a_{21} & a_{22} & a_{23} \\ a_{11} & a_{12} & a_{13} \end{pmatrix}.$$

We see that  $E$  has the required effect on *any*  $3 \times 3$  matrix. This leads us to the following definition.

A matrix obtained from the unit matrix by an elementary row operation is called an **elementary matrix**.

### Example 2

We shall find the elementary matrices corresponding to the row operations

$$E_1: R_1 \longleftrightarrow R_2, E_2: R_2 \longmapsto 3R_2 \text{ and } E_3: R_2 \longmapsto R_2 - 5R_3$$

which we used in Example 1. We shall then *premultiply* the matrix  $A$  of Example 1 by the elementary matrices found, and note the results obtained.



We begin with the identity matrix:

$$I = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$E_1: R_1 \longleftrightarrow R_2$$

Writing  $E_1$  to stand for the matrix as well as the operation, we get

$$E_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix};$$

$$E_1 A = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 3 \\ 1 & 1 & 1 \\ 1 & 0 & -2 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 3 \\ 1 & 0 & -2 \end{pmatrix}.$$

$$E_2: R_2 \longmapsto 3R_2$$

$$E_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{pmatrix};$$

$$E_2 A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 3 \\ 1 & 1 & 1 \\ 1 & 0 & -2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 3 \\ 3 & 3 & 3 \\ 1 & 0 & -2 \end{pmatrix}.$$

$$E_3: R_2 \longmapsto R_2 - 5R_3$$

$$E_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -5 \\ 0 & 0 & 1 \end{pmatrix};$$

$$E_3 A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -5 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 3 \\ 1 & 1 & 1 \\ 1 & 0 & -2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 3 \\ -4 & 1 & 11 \\ 1 & 0 & -2 \end{pmatrix}$$

You should compare these results with those of Example 1.



*Exercise 1*

Find the elementary matrices of order  $3 \times 3$  corresponding to each of the following:

- (i)  $R_2 \longleftrightarrow R_3$
- (ii)  $R_2 \longmapsto R_2 - 2R_1$
- (iii)  $R_3 \longmapsto R_3 + 2R_1 + 3R_2$
- (iv)  $R_2 \longmapsto R_2 - R_1 + 2R_3$

Verify that the elementary matrices have the desired effect by applying them to the matrix

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ -1 & -1 & -2 \end{pmatrix}$$

We have seen that, to each elementary row operation used in the Gauss-elimination method, there corresponds an elementary matrix.

It appears that we have an isomorphism between

- (1) the set of elementary operations carried out on the system of simultaneous equations, combined by successive performance, and
- (2) the set of elementary matrices, combined by matrix multiplication.

The object of the Gauss elimination method is to use a finite sequence of elementary operations to reduce a system of equations into any *equivalent* system which can be solved simply by back-substitution. We shall now reinterpret this in terms of matrices.

Suppose we start with the system of equations written in matrix form as

$$A\underline{x} = \underline{b},$$

that is,

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}.$$

The object is to reduce this system to the equivalent system

$$C\underline{x} = \underline{d},$$



that is,

$$\begin{pmatrix} c_{11} & c_{12} & c_{13} \\ 0 & c_{22} & c_{23} \\ 0 & 0 & c_{33} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ d_3 \end{pmatrix}.$$

If  $E_1$  represents the matrix corresponding to the first elementary operation which we carry out on the original system, we now premultiply both sides of the matrix operation by  $E_1$ , to obtain

$$E_1 A \underline{x} = E_1 \underline{b}.$$

This new equation represents a system of equations equivalent to the first. Notice that by premultiplying  $A$  and  $\underline{b}$  by  $E_1$  we are effectively doing the elementary operation. So there is no point, in any practical calculation, in finding the matrix  $E$  and actually doing the matrix multiplication: it is much easier to carry out the appropriate row operation on the augmented matrix  $(A \ \underline{b})$ . The *only* value in knowing of the existence of the matrix  $E$  is in theoretical considerations which may have practical consequences, but the matrix  $E$  itself is not used in practice.

We have seen that each elementary operation corresponds to an elementary matrix. It follows that a sequence of elementary operations corresponds to a product of elementary matrices. Suppose the sequence of matrices, in order of usage, is  $E_1, E_2, E_3, \dots, E_s$ . Then, what we are doing can be written in the form

$$(E_s \dots E_3 E_2 E_1 A) \underline{x} = E_s \dots E_3 E_2 E_1 \underline{b}.$$

Notice that, because we are always premultiplying by the  $E$ 's, the column vector  $\underline{x}$  is never involved. When setting out numerical calculations we drop the  $\underline{x}$  and just keep the augmented matrix  $(A \ \underline{b})$  and manipulate that. We have

$$\begin{aligned} (C \ \underline{d}) &= (E_s \dots E_3 E_2 E_1)(A \ \underline{b}) \\ &= P(A \ \underline{b}), \end{aligned}$$

where  $P$  is the matrix obtained by multiplying all the  $E$ 's together. Incidentally, we know that *if* we can get from  $(A \ \underline{b})$  to  $(C \ \underline{d})$ , we can reverse each step (each elementary operation can be reversed by another elementary operation of the same kind) and get from  $(C \ \underline{d})$  back to  $(A \ \underline{b})$ . This means that the matrix  $P$  is non-singular, i.e. it has an inverse matrix  $P^{-1}$ . Notice that we said "*if* we can get from  $(A \ \underline{b})$  to  $(C \ \underline{d})$ ": we have no guarantee that we can. In fact, it is always possible, although some of the



$c$ 's may be zero. If the solution of the system of equations is unique, i.e. if  $r(A) = 3$  (or, in general, if  $r(A) = n$  for a square  $n \times n$  matrix), then none of the  $c$ 's on the leading diagonal is zero.

We have now achieved the object of this section, which was to interpret the Gauss elimination method in terms of matrices. There are a number of interesting consequences and refinements which we shall consider in subsequent sections.

### Exercise 2

Find elementary matrices which reduce the system

$$\begin{pmatrix} 2 & 1 & -1 \\ 6 & -1 & -9 \\ 4 & 3 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 3 \\ 7 \\ 5 \end{pmatrix}$$

to the system

$$\begin{pmatrix} 2 & 1 & -1 \\ 0 & -4 & -6 \\ 0 & 0 & \frac{3}{2} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 3 \\ -2 \\ -\frac{3}{2} \end{pmatrix}.$$

Hence find the matrix  $P$  and verify that

$$\left( \begin{array}{ccc|c} 2 & 1 & -1 & 3 \\ 0 & -4 & -6 & -2 \\ 0 & 0 & \frac{3}{2} & -\frac{3}{2} \end{array} \right) = P \left( \begin{array}{ccc|c} 2 & 1 & -1 & 3 \\ 6 & -1 & -9 & 7 \\ 4 & 3 & 1 & 5 \end{array} \right)$$

## 8.2 The Inverse of a Matrix

There are a number of ways of finding the inverse of a non-singular matrix  $A$ . In this section we shall exploit the elementary matrix technique discussed in the last section. This is a good example of the way in which elementary matrices, although not themselves practical, lead to practical methods.

Suppose that  $A$  is a non-singular matrix and that  $X$  is its inverse. Then

$$XA = AX = I,$$

where  $A$ ,  $X$  and  $I$  are square matrices of the same order. We can consider the equation  $AX = I$  as an equation from which we wish to determine



the unknown matrix  $X$ . It is, in fact, not very different from our previous equation. For instance, if we suppose that  $A$ ,  $X$  and  $I$  are all  $3 \times 3$ , then we can write

$$A(\underline{x}_1 \quad \underline{x}_2 \quad \underline{x}_3) = (\underline{i}_1 \quad \underline{i}_2 \quad \underline{i}_3),$$

where we have expressed the matrices  $X$  and  $I$  in terms of their column vectors in the usual way. This *one* matrix equation is then equivalent to the *three* matrix equations

$$A\underline{x}_1 = \underline{i}_1, \quad A\underline{x}_2 = \underline{i}_2, \quad A\underline{x}_3 = \underline{i}_3,$$

and we are back to the problem of solving the equation  $A\underline{x} = \underline{b}$ , except that we now have to solve three matrix equations instead of one.

This suggests that the same approach might help here. We can perhaps find a sequence of elementary operations (with corresponding elementary matrices) which together form the Gauss elimination method. The same sequence of operations would, of course, do for all the three equations. This approach would certainly work, but a little more effort will give a much bigger return. The Gauss elimination method requires back-substitution; we can avoid back-substitution if we carry out some more elementary operations. Instead of reducing  $A$  (in the  $3 \times 3$  case) to the form

$$\begin{pmatrix} c_{11} & c_{12} & c_{13} \\ 0 & c_{22} & c_{23} \\ 0 & 0 & c_{33} \end{pmatrix},$$

we could carry on, and try to reduce it all the way down to

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

still using elementary operations. Then we could read off the solution without back-substitution.

Let us suppose that we can find a sequence of elementary operations which transform  $A$  into  $I$ . We shall denote the corresponding elementary matrices by  $E_1, E_2, \dots, E_s$ . Then starting from our original equation

$$AX = I,$$



we have

$$(E_s \dots E_2 E_1 A)X = E_s \dots E_2 E_1 I,$$

that is,

$$IX = E_s \dots E_2 E_1 I,$$

or

$$X = E_s \dots E_2 E_1 I.$$

This is a remarkable result which has practical consequences. It tells us that if we can find a sequence of elementary operations which transforms  $A$  into  $I$ , that same sequence of operations will transform  $I$  into the inverse of  $A$ . The elementary matrices give us the justification, but in practice we do not use these—we use their effects—the elementary operations.

### Example 1

Find the inverse of the matrix

$$A = \begin{pmatrix} 2 & 1 & -1 \\ 6 & -1 & -9 \\ 4 & 3 & 1 \end{pmatrix}.$$

There is no harm in assuming that the inverse exists: we have no reason to suppose either that it does or that it does not, but the proof of the pudding will be in the eating. If we can find a sequence of elementary operations which transforms  $A$  into  $I$ , then the inverse of  $A$  exists. Since we are going to perform the same elementary operations on the rows of  $A$  and  $I$ , we get ourselves into battle array, dropping matrix brackets.

$$\begin{array}{ccc|ccc|c} 2 & 1 & -1 & 1 & 0 & 0 & 3 \\ 6 & -1 & -9 & 0 & 1 & 0 & -3 \\ 4 & 3 & 1 & 0 & 0 & 1 & 9 \end{array}$$

You may be wondering where the last column came from. As we mentioned in our notes on the Gauss elimination method, any good numerical method should incorporate a check of some sort. So we have introduced an extra figure at the end of each row, which is the sum of the numbers in that row. In the calculation we treat it as part of that row, so that whatever we do to the row (by an elementary operation), the last number should still be the sum of the numbers in that row. If, after a certain step, we find that the sum-check fails, then there is an error somewhere in that step. As long as we remember to check the row sum after each step, we



should be *reasonably* sure of getting the whole calculation right. (Only *reasonably* sure, because we may have made compensating errors.) We now proceed *systematically* to produce the matrix  $I$  in the first three columns, indicating each step by our usual notation.

$$R_1 \longmapsto \frac{1}{2}R_1$$

$$\begin{array}{ccc|ccc|c} 1 & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & 0 & 0 & \frac{3}{2} \\ 6 & -1 & -9 & 0 & 1 & 0 & -3 \\ 4 & 3 & 1 & 0 & 0 & 1 & 9 \end{array}$$

(When you get adept at the game, then you don't need to copy down *all* the rows at each step, but only the ones that change; for example, the first row above. We shall not do this because it can be a bit confusing for the beginner, but we shall occasionally do more than one step at one go.)

$$R_2 \longmapsto R_2 - 6R_1; R_3 \longmapsto R_3 - 4R_1$$

$$\begin{array}{ccc|ccc|c} 1 & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & 0 & 0 & \frac{3}{2} \\ 0 & -4 & -6 & -3 & 1 & 0 & -12 \\ 0 & 1 & 3 & -2 & 0 & 1 & 3 \end{array}$$

We now have one column correct.

$$R_2 \longmapsto -\frac{1}{4}R_2$$

$$\begin{array}{ccc|ccc|c} 1 & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & 0 & 0 & \frac{3}{2} \\ 0 & 1 & \frac{3}{2} & \frac{3}{4} & -\frac{1}{4} & 0 & 3 \\ 0 & 1 & 3 & -2 & 0 & 1 & 3 \end{array}$$

$$R_1 \longmapsto R_1 - \frac{1}{2}R_2; R_3 \longmapsto R_3 - R_2$$

$$\begin{array}{ccc|ccc|c} 1 & 0 & -\frac{5}{4} & \frac{1}{8} & \frac{1}{8} & 0 & 0 \\ 0 & 1 & \frac{3}{2} & \frac{3}{4} & -\frac{1}{4} & 0 & 3 \\ 0 & 0 & \frac{3}{2} & -\frac{11}{4} & \frac{1}{4} & 1 & 0 \end{array}$$

We now have two columns correct.

$$R_3 \longmapsto \frac{2}{3}R_3$$

$$\begin{array}{ccc|ccc|c} 1 & 0 & -\frac{5}{4} & \frac{1}{8} & \frac{1}{8} & 0 & 0 \\ 0 & 1 & \frac{3}{2} & \frac{3}{4} & -\frac{1}{4} & 0 & 3 \\ 0 & 0 & 1 & -\frac{11}{6} & \frac{1}{6} & \frac{2}{3} & 0 \end{array}$$



$$R_1 \mapsto R_1 + \frac{5}{4}R_3; R_2 \mapsto R_2 - \frac{3}{2}R_3$$

$$\begin{array}{ccc|ccc|c} 1 & 0 & 0 & -\frac{13}{6} & \frac{1}{3} & \frac{5}{6} & 0 \\ 0 & 1 & 0 & \frac{7}{2} & -\frac{1}{2} & -1 & 3 \\ 0 & 0 & 1 & -\frac{11}{6} & \frac{1}{6} & \frac{2}{3} & 0 \end{array}$$

We now have all three columns correct.

If our theory is correct, then the inverse of  $A$  is

$$\begin{pmatrix} -\frac{13}{6} & \frac{1}{3} & \frac{5}{6} \\ \frac{7}{2} & -\frac{1}{2} & -1 \\ -\frac{11}{6} & \frac{1}{6} & \frac{2}{3} \end{pmatrix}$$

To check that this matrix is indeed the inverse of the matrix  $A$ , do Exercise 1.

### Exercise 1

Given that

$$A = \begin{pmatrix} 2 & 1 & -1 \\ 6 & -1 & -9 \\ 4 & 3 & 1 \end{pmatrix} \text{ and } X = \begin{pmatrix} -\frac{13}{6} & \frac{1}{3} & \frac{5}{6} \\ \frac{7}{2} & -\frac{1}{2} & -1 \\ -\frac{11}{6} & \frac{1}{6} & \frac{2}{3} \end{pmatrix},$$

show that

$$XA = I = AX,$$

where

$$I = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

### Exercise 2

Find the inverse matrices of

$$(i) \begin{pmatrix} 1 & -1 & 0 \\ 3 & 7 & 1 \\ 2 & -14 & 2 \end{pmatrix} \quad (ii) \begin{pmatrix} 0 & 1 & -1 \\ 3 & 6 & 2 \\ -1 & 0 & 5 \end{pmatrix}.$$



We now have a technique for finding the inverse of any non-singular matrix. The question may naturally arise in your mind: What is the use of it? We have briefly answered this question earlier, but we shall now discuss it more fully.

If we have a system of equations

$$A\underline{x} = \underline{b},$$

where  $A$  is a non-singular square matrix, then, since the inverse matrix,  $A^{-1}$ , represents the inverse isomorphism, there is a unique solution to the system; it is

$$\underline{x} = A^{-1}\underline{b}.$$

The important thing about it is that, once we know  $A^{-1}$ , we can find a unique solution for *any* column vector  $\underline{b}$ . This is a distinct advantage over the Gauss elimination method, where, for a different choice of  $\underline{b}$ , we would have to do the Gauss elimination again. (Although, of course, only the  $\underline{b}$  column would be different and so, provided we had kept a record of our previous calculations, the work involved would not be as hard as it may seem at first sight.)

Although we have restricted our considerations here to non-singular matrices, and we have no effective method of looking at a square matrix and telling whether it is non-singular, the method described in this section is still useful. Even if it does not lead to the inverse matrix, some of the same steps will lead to a solution (or solutions) if one exists: if no solution exists, it will tell us this as well. We shall not go into these points here. They are not difficult, but we have gone far enough into the theory of linear equations for the time being.

We shall turn briefly, in the next section, to the calculation of rank. This involves a similar discussion to the one above.

### 8.3 Calculation of the Rank of a Matrix

Although we have stated some of the results in Chapter 7 in terms of the rank of a matrix, we have not discussed an effective way of calculating it (our examples were artificially easy). In this section we shall make good this deficiency. We shall restrict our discussion to square matrices for simplicity.

Given a square matrix  $A$ , we have defined the rank of  $A$  as the maximum number of linearly independent column vectors in  $A$ .



If  $A$  is an  $n \times n$  matrix, then we can associate with  $A$  the mapping

$$A:\underline{x} \longmapsto A\underline{x} \quad (\underline{x} \in R^n),$$

and we have shown in section 7.6 that

$$r(A) = \text{dimension of } A(R^n).$$

We shall consider the effect of one of the elementary row operations on the rank of  $A$ . Although we cannot necessarily readily tell the rank of  $A$ , we can tell the rank of a much simplified associated matrix (for instance, of the form obtained in the Gauss elimination process).

Let  $E$  be the matrix corresponding to one of the elementary row operations. The corresponding mapping

$$E:\underline{x} \longmapsto E\underline{x} \quad (\underline{x} \in R^n)$$

is one-one, because, as we have already noted, every elementary row operation has an inverse operation of the same type.

Consider the composite mapping

$$E \circ A:\underline{x} \longmapsto EA\underline{x} \quad (x \in R^n)$$

with matrix  $EA$ . We have

$$\begin{aligned} r(EA) &= \text{dimension of } E \circ A(R^n) \\ &= \text{dimension of } E(A(R^n)) \end{aligned}$$

But, as we saw in section 7.6, since  $E$  is one-one, the kernel of the  $E$  mapping contains just the zero element, and so is of dimension zero; that is,

$$\text{dimension of } E(A(R^n)) = \text{dimension of } A(R^n) = r(A).$$

So

$$r(A) = r(EA).$$

This means that the rank of  $A$  is *unaffected by an elementary row operation*. This leads to a practical method of calculating rank.

### Example 1

We calculate the rank of the matrix

$$A = \begin{pmatrix} 1 & 0 & -1 \\ -2 & 1 & 3 \\ 4 & 2 & -2 \end{pmatrix}$$



by *systematically* reducing the elements below the leading diagonal (from top left to bottom right) to zero, using elementary row operations. We use a row sum, as described in the previous section, to act as a check for the arithmetic.

$$\begin{array}{ccc|c} 1 & 0 & -1 & 0 \\ -2 & 1 & 3 & 2 \\ 4 & 2 & -2 & 4 \end{array}$$

$$R_2 \longmapsto R_2 + 2R_1; R_3 \longmapsto R_3 - 4R_1$$

$$\begin{array}{ccc|c} 1 & 0 & -1 & 0 \\ 0 & 1 & 1 & 2 \\ 0 & 2 & 2 & 4 \end{array}$$

$$R_3 \longmapsto R_3 - 2R_2$$

$$\begin{array}{ccc|c} 1 & 0 & -1 & 0 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{array}$$

It is now easy to see that the first two columns are linearly independent, but

$$\begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} = -1 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix},$$

so  $r(A) = 2$ .

In fact, if we use the result, mentioned in section 7.5, that the row rank (the maximum number of linearly independent row vectors) is equal to the column rank, we can see the rank of  $A$  even more easily, since the last row consists entirely of zeros. In general, if we carry out this reduction, the rank of  $A$  is equal to the number of “non-zero” rows in the reduced matrix.



*Exercise 1*

Calculate the rank of the following matrix:

$$\begin{pmatrix} 1 & 2 & 1 \\ 0 & 1 & 1 \\ 2 & -1 & -1 \end{pmatrix}.$$

*Exercise 2*

Use methods similar to those in the text to prove that, if  $B$  is an  $n \times n$  non-singular matrix and  $A$  is any  $n \times n$  matrix, then

$$r(BA) = r(A).$$

**Summary**

In the first section of this chapter we defined elementary matrices. These correspond to the elementary operations in the Gauss elimination method. We thus obtained a matrix representation of this method. Although this had no direct practical interest, it had practical consequences. It allowed us to develop:

- (i) a method for finding the inverse of a non-singular matrix, and
- (ii) a method for calculating the rank.

Since both these methods use essentially the same systematic procedure as the Gauss elimination method, we can use the same calculations to determine:

- (i) the solution to a system of equations,
  - (ii) the inverse of the matrix of coefficients, if it exists,
  - (iii) the rank of the matrix of coefficients or the augmented matrix,
- whichever is of interest.

**8.4 Direct Methods**

In the remainder of this chapter we look at various methods of solving systems of linear equations. We shall be particularly interested in the *efficiency* of these methods, and we shall, therefore, include in our discussion the Gauss elimination method. We start with the so-called *direct method*.



## Introduction

Mathematicians like to be able to record an explicit solution to an equation. The oft-quoted solution to the quadratic equation

$$ax^2 + bx + c = 0,$$

namely

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a},$$

is an example of this. The desired variable is to the left of an “equals” sign, and the letters to the right of the “equals” sign can be replaced by numbers from the data in any specific example. If  $b^2 > 4ac$  the solution consists of the two numbers which map to zero under the function

$$x \longmapsto ax^2 + bx + c \quad (x \in R).$$

Similarly, we can write down an explicit solution to the matrix equation

$$A\underline{x} = \underline{b},$$

where  $A$  is non-singular, in the form

$$\underline{x} = A^{-1}\underline{b},$$

the solution vector being the unique vector which maps to  $\underline{b}$  under the mapping

$$\underline{x} \longmapsto A\underline{x} \quad (\underline{x} \in R^n).$$

A method which uses such an explicit formula to obtain the answer is called a **direct method**. Not all direct methods, however, use formulas to calculate the answer. For instance, the Gauss elimination method is a direct method, but we do not *use* a formula for the answer, but a step by step procedure to get from the data to the answer. This is the characteristic of a direct method (as opposed to an iterative method, which we discuss later): it is a method of obtaining the answer to a problem from the data by a step by step procedure, the answer (or an approximation to the answer) being obtained only at the end of the procedure, there being no approximations to the answer en route.

To *calculate* the solution vector to a system of equations using the formula  $\underline{x} = A^{-1}\underline{b}$ , we must determine  $A^{-1}$  and post-multiply by  $\underline{b}$ . We examine the efficiency of this particular process in this section. We shall also compare its efficiency with the efficiency of two other methods: the Gauss elimination method and one other direct method. Finally, we must



emphasize that all three methods of solution are algebraically equivalent, since the solution is unique. This means that if we wrote them out as explicit formulas, we could derive one formula from the other by algebraic manipulation. In this section we are interested in comparing the computational methods on the basis of the number of arithmetic operations required to apply them.

### An Explicit Method

By appropriate manipulation (for example, eliminating  $x_2$  between the equations, and then solving for  $x_1$ ), the solution of

$$a_{11}x_1 + a_{12}x_2 = b_1$$

$$a_{21}x_1 + a_{22}x_2 = b_2$$

may be written as

$$x_1 = \frac{b_1a_{22} - b_2a_{12}}{a_{11}a_{22} - a_{12}a_{21}}, \quad x_2 = \frac{a_{11}b_2 - a_{21}b_1}{a_{11}a_{22} - a_{12}a_{21}},$$

provided that

$$a_{11}a_{22} - a_{12}a_{21} \neq 0.$$

In this case the solution we have found is the set of components of the unique vector  $\underline{x}$  which maps to  $\underline{b}$  under the mapping

$$\underline{x} \longmapsto A\underline{x} \quad (\underline{x} \in R^2).$$

Thus, at one step, provided the condition is satisfied, we have formally written down the solution to *all* pairs of simultaneous equations in two variables, which have a unique solution.

We can go on to find that for a system of three simultaneous linear equations:

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1$$

$$a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = b_2$$

$$a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = b_3$$

the solution set may be written as

$$x_1 = \frac{b_1(a_{22}a_{33} - a_{23}a_{32}) - b_2(a_{12}a_{33} - a_{13}a_{32}) + b_3(a_{12}a_{23} - a_{13}a_{22})}{a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{21}(a_{12}a_{33} - a_{13}a_{32}) + a_{31}(a_{12}a_{23} - a_{13}a_{22})},$$

together with similar expressions for  $x_2$  and  $x_3$  (with the same denominator), provided, of course, that the denominator does not equal zero. (If



you have plenty of time to spare, you may like to verify the above expression by solving the system of equations yourself. But don't get too bogged down by the algebraic manipulations.)

Again, we can see that, by simple substitution of numbers for the letters, we can solve *any* system of the above form which has a unique solution. We could, in theory, solve a system of simultaneous linear equations of any order by developing the appropriate formulas. Why then do we not stop here and simply solve simultaneous linear equations this way? The reason is that the expenditure of time and effort involved is too great.

Let us consider the number of multiplications and divisions involved in solving the three simultaneous linear equations. We concentrate on the operations of multiplication and division, since these tend to be more time-consuming than addition and subtraction, both for manual and computer operations.

The expressions for  $x_1$ ,  $x_2$  and  $x_3$  have the same denominator, so this has to be calculated just once. Each numerator has the same form as the denominator, and there are 3 numerators. Thus we need to calculate 4 expressions of the form:

$$a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{21}(a_{12}a_{33} - a_{13}a_{32}) + a_{31}(a_{12}a_{23} - a_{13}a_{22}).$$

Each bracket contains 2 products. Evaluation of the whole expression therefore requires 9 multiplications.

$$\text{Total number of multiplications is } 4 \times 9 = 36$$

$$\text{Number of divisions} = 3$$

$$\text{Therefore total number of time-consuming operations} = 39$$

You may have noticed that three bracketed terms in one numerator (that of  $x_1$  in our example) are the same as the bracketed terms in the denominator. Using this, we could save 6 multiplications and would therefore require only 33 time-consuming operations.

### *Exercise 1*

What is the total number of operations of multiplication and division required to solve a system of two simultaneous linear equations by substitution in the formulas given in the text?

A general expression can be derived for the number of operations of multiplication and division for the solution of  $n$  equations in  $n$  unknowns, by this method; it is given by the sequence

$$u_2 = 8$$



$$u_n = 2n + n^2\left(\frac{u_{n-1}}{n-1} - 1\right), \quad n > 2.$$

For large values of  $n$ , the  $n$ th term of this sequence,  $u_n$ , can be shown to be approximately equal to

$$1.72 \times n \times n!$$

This enables us to produce the table below. The final column gives a rough idea of the time it would take an automatic computer to solve the problem if, for example, each operation were to take one microsecond (1 microsecond =  $10^{-6}$  s; it is denoted by 1  $\mu$ s).

Number of simultaneous equations	Number of operations	Time taken
2	8	8 $\mu$ s
3	33	33 $\mu$ s
4	168	168 $\mu$ s
10	$\simeq 6.2 \times 10^7$	62 s
100	$\simeq 1.6 \times 10^{160}$	$\simeq 10^{147}$ year
1000	$\simeq 6.9 \times 10^{2570}$	$\simeq 10^{2557}$ year

This shows the tremendous increase in the number of operations and the consequent time required as  $n$  increases; a more efficient method is obviously desirable.

Determinants

If you wish, you may omit this section, which is not essential to the development. It is included because it briefly discusses *determinants*, which have in the past been closely associated with the solution of simultaneous linear equations.

The expression

$$a_{11}a_{22} - a_{12}a_{21}$$



is called the *determinant* of the matrix

$$A_2 = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix},$$

and is written as

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}$$

or  $\det A_2$  or  $|A_2|$ , so that the solution of two equations in two variables may be written as

$$x_1 = \frac{\begin{vmatrix} b_1 & a_{12} \\ b_2 & a_{22} \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}} \text{ etc.}$$

Similarly, the expression

$$a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{21}(a_{12}a_{33} - a_{13}a_{32}) + a_{31}(a_{12}a_{23} - a_{13}a_{22})$$

is called the *determinant* of the matrix

$$A_3 = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix},$$

and is written as

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}$$

or  $\det A_3$  or  $|A_3|$ , so that the solution of three equations in three unknowns may be written as

$$x_1 = \frac{\begin{vmatrix} b_1 & a_{12} & a_{13} \\ b_2 & a_{22} & a_{23} \\ b_3 & a_{32} & a_{33} \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}}.$$

etc.



The determinant of a square matrix is a certain number associated with the matrix. Essentially, its use at this level in mathematics is confined to being a shorthand for expressing formally the solutions of simultaneous linear equations. Since actual calculations using the formula for the solution, which are equivalent to evaluating the determinants, are very time-consuming, as we shall see below, we have not made the discussion of the properties of determinants an integral part of this volume.

### The Gauss Elimination Method

In Chapter 7, we introduced the Gauss elimination method of solving systems of equations. The objective of the method is to produce an equivalent set of equations which are easier to solve than the original set. We now ask you to consider this method again and to check how efficient it is in terms of the number of operations it takes to produce the answer, since for large systems of equations the method discussed at the beginning of this section is clearly unacceptable. (Even if somebody had started a modern computer solving 100 simultaneous equations by that method at the beginning of this century, it would still not have made a dent in the problem: it would not even have performed  $10^{-100}\%$  of the calculations.)

#### Exercise 2

Solve the following set of three equations by the Gauss elimination method with back substitution, in the order shown in the scheme below. The numbers in your calculation should be expressed as decimals, not fractions. Note the number of divisions and multiplications that have occurred. What is the total number of these operations for the whole solution?

$$5x_1 + 2x_2 - 2x_3 = 8 \quad (1)$$

$$3x_1 + 6x_2 - 4x_3 = 1 \quad (2)$$

$$2x_1 + 4x_2 - 2x_3 = 1 \quad (3)$$



Step in the Calculation	Number of Multiplications and/or Divisions
<p>Eliminate <math>x_1</math> from equations (2) and (3).</p> <p>Factor for equation (2) is <input type="text"/></p> <p>Factor for equation (3) is <input type="text"/></p> $5x_1 + 2x_2 + (-2) \quad x_3 = 8 \quad (1)$ $\text{} x_2 + \text{} x_3 = \text{} \quad (4)$ $\text{} x_2 + \text{} x_3 = \text{} \quad (5)$ <p>Now eliminate <math>x_2</math> from equation (5).</p> <p>Factor for equation (5) is <input type="text"/></p> $5x_1 + 2x_2 + (-2) \quad x_3 = 8 \quad (1)$ $\text{} x_2 + \text{} x_3 = \text{} \quad (4)$ $\text{} x_3 = \text{} \quad (6)$ <p>Therefore</p> $x_3 = \text{}$ <p>and back-substituting into equation (4) gives</p> $x_2 = \text{}$ <p>and from equation (1)</p> $x_1 = \text{}$	
Total number of divisions and multiplications	



Next we try to determine the amount of computation required to solve a general system of equations by the Gauss elimination method. We attempt an analysis similar to the one given in the last exercise. We then compare the labour involved in this method with the labour involved in the method at the beginning of this section.

Consider the first two equations of a system of  $n$  equations:

$$a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n = b_1$$

$$a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n = b_2$$

We assume that  $a_{11}$  is non-zero. If it were not, then we could alter the order of the equations until we found a coefficient of  $x_1$  which was non-zero. In hand calculation, such a re-ordering does not involve an important time factor, but in an automatic machine calculation it would, of course, have to be taken into account. To eliminate  $x_1$  from the second equation requires the calculation of the quotient  $a_{21}/a_{11}$  and then the formation of the  $n$  numbers

$$a_{22} - \frac{a_{21}}{a_{11}}a_{12}, a_{23} - \frac{a_{21}}{a_{11}}a_{13}, \dots, a_{2n} - \frac{a_{21}}{a_{11}}a_{1n}, b_2 - \frac{a_{21}}{a_{11}}b_1.$$

Thus the formation of the new second equation requires  $(n + 1)$  operations of multiplication or division. In the next exercise we ask you to calculate the total number of multiplications and divisions required.

We assume that  $a_{22} - \frac{a_{21}}{a_{11}}a_{12}$  is non-zero for the next round, since we

shall have to divide by it to produce the next quotient corresponding to

$\frac{a_{21}}{a_{11}}$ , that is,

$$\frac{a_{32}}{a_{22} - \frac{a_{21}}{a_{11}}a_{12}}.$$

If it were zero, we could alter the order of the equations until we found a coefficient of  $x_2$  which was non-zero. (What would you infer if *all* subsequent coefficients of  $x_2$  were zero?)

In general, we shall assume that we can perform the divisions as we come to them.



Exercise 3

Fill in the table on p. 291 to calculate the number of multiplications and divisions involved in solving a system of  $n$  equations in  $n$  unknowns by the Gauss elimination method. The formulas

$$1 + 2 + 3 + \cdots + n = \sum_{r=1}^n r = \frac{n(n + 1)}{2}$$
$$1^2 + 2^2 + 3^2 + \cdots + n^2 = \sum_{r=1}^n r^2 = \frac{n(n + 1)(2n + 1)}{6},$$

will be useful.

We are now in a position to compare the explicit solution and the Gauss elimination method from the viewpoint of efficiency. The results are tabulated below.

Number of equations	Formula method No. of operations is $1.72n \times n!$ (for large $n$ )	Gauss elimination method No. of operations is $\frac{n^3}{3} + n^2 - \frac{n}{3}$
3	33	17
4	168	36
10	$\simeq 6.2 \times 10^7$	430
100	$\simeq 1.6 \times 10^{160}$	$\simeq 3.4 \times 10^5$
1000	$\simeq 6.9 \times 10^{2570}$	$\simeq 3.3 \times 10^8$

This table demonstrates dramatically why the simple formula method is never used for computing numerical solutions for systems of linear equations.

In fact, the formula method is never quicker than the Gauss elimination method. For a quick comparison of the methods for large  $n$ , we would say that the number of operations for the Gauss elimination method is approximately  $\frac{n^3}{3}$ , since, when  $n > 100$  say, the rest of the terms affect the number of operations by less than three per cent.



Step in the Calculation	Number of Multiplications/Divisions
Eliminate $x_1$ from all equations after the first.	<div>(n - )(n + 1) = n<sup>2</sup> - 1</div>
Eliminate $x_2$ from all equations after the second. ⋮	<div>=</div>
Eliminate $x_{n-1}$ from all equations after the $(n - 1)$ th, i.e. the $n$ th equation.	<div></div>
Total for elimination is	<div></div>
	<div>=</div>
Now carry out the back substitution.	
To determine $x_n$ from an equation such as $\alpha_n x_n = \beta_n$	<div></div>
To determine $x_{n-1}$ from an equation such as $\gamma_{n-1} x_{n-1} + \gamma_n x_n = \beta_{n-1}$ ⋮	<div></div>
To determine $x_1$	<div></div>
Total for back substitution is	<div></div>
Therefore total number of operations is	<div></div>

Compare your answers with the answer to Exercise 1.



### A Matrix Inversion Method

In section 8.2 we described a numerical method for finding the inverse of an  $n \times n$  matrix; the method is very similar to the Gauss elimination method. We pointed out that if we had to solve several systems of equations of the form

$$A\underline{x} = \underline{b},$$

in which the matrix  $A$  was the same in each case, and only the matrix  $\underline{b}$  changed, then there might be some value in computing  $A^{-1}$  and calculating the solution from the formula

$$\underline{x} = A^{-1}\underline{b}$$

for each case.

We shall now look at the number of calculations involved and see just when this method is of "some value".

Just to remind you of the method, we describe the essential steps. Suppose we want to invert the matrix

$$\begin{pmatrix} 2 & 1 & -1 \\ 6 & -1 & -9 \\ 4 & 3 & 1 \end{pmatrix}.$$

We write

$$\begin{array}{ccc|ccc} 2 & 1 & -1 & 1 & 0 & 0 \\ 6 & -1 & -9 & 0 & 1 & 0 \\ 4 & 3 & 1 & 0 & 0 & 1 \end{array}$$

and use elementary row operations to turn this array into

$$\begin{array}{ccc|ccc} 1 & 0 & 0 & \times & \times & \times \\ 0 & 1 & 0 & \times & \times & \times \\ 0 & 0 & 1 & \times & \times & \times \end{array}$$

where the crosses represent the elements of the inverse matrix. We have ignored the row-sum check: this would be involved both here and in the Gauss elimination method, so for a comparison between the two methods we can ignore it.

We begin by dividing the first row of our array of numbers by 2 in order to obtain a 1 in the top left-hand corner. This involves 3 divisions. We obtain

$$\begin{array}{ccc|ccc} 1 & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 6 & -1 & -9 & 0 & 1 & 0 \\ 4 & 3 & 1 & 0 & 0 & 1 \end{array}$$



Next we subtract 6 times the first row from the second and 4 times the first row from the third. This involves  $2 \times 3$  multiplications. We obtain

$$\begin{array}{ccc|ccc} 1 & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & -4 & -6 & -3 & 1 & 0 \\ 0 & 1 & 3 & -2 & 0 & 1 \end{array}$$

Thus to get the first column into the required form we have performed 9 time-consuming operations.

In general, to compute the inverse of an  $n \times n$  matrix, we form the corresponding  $n \times 2n$  matrix by “joining” on the  $n \times n$  unit matrix.

To produce

$$\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

in the first column requires  $n$  divisions to produce the 1 and  $(n - 1) \times n$  operations to produce the 0's. That means we require  $n + (n - 1) \times n = n^2$  time-consuming operations to derive the first column. Performing the manipulations to produce each subsequent column of an  $n \times n$  unit matrix also requires  $n^2$  operations. Since there are  $n$  columns, the total number of operations required is  $n^3$ .

Having found the inverse matrix, we then have to calculate  $A^{-1}\underline{b}$ , which involves a further  $n^2$  multiplications.

Thus the total number of operations to obtain a solution is

$$n^3 + n^2,$$

compared with

$$\frac{n^3}{3} + n^2 - \frac{n}{3}$$

for the Gauss elimination method.

Number of equations	Gauss elimination No. of operations	Inverse method No. of operations
3	17	36
4	36	80
10	430	1100
100	$\simeq 3.4 \times 10^5$	$\simeq 3 \times 3.4 \times 10^5$
1000	$\simeq 3.3 \times 10^8$	$\simeq 3 \times 3.3 \times 10^8$



For large  $n$ , the  $n^3$  term is much larger than the rest, so that the inverse method is roughly three times as long as the Gauss elimination method.

Exercise 4

How many multiplications are required to multiply together two  $n \times n$  matrices?

Summary

In this section we have measured the efficiency, in terms of time for computation, of three direct methods of solving systems of  $n$  equations (in  $n$  unknowns) by investigating the number of operations of multiplication and division required for each. For large  $n$ , we found the following results.

	Approximate number of operations	Usefulness for large $n$
Explicit method	$1.72n \times n!$	Of no use for computing solutions
Gauss elimination method.	$\frac{n^3}{3}$	The most economical direct method of solving systems of linear equations.
A method using the inverse of a matrix.	$n^3$	Computationally less efficient than Gauss elimination.

The methods considered in this section are basic methods. There are many refinements to them for particular problems, some of which are listed, for example, in L. Fox, *An Introduction to Numerical Linear Algebra* (Clarendon Press, 1964), Chapters 3 and 4.



## 8.5 Iterative or Indirect Methods

### Introduction

Theoretically, given enough time and ignoring round-off errors which may occur in the computation, it is possible to solve exactly a matrix equation

$$A\underline{x} = \underline{b},$$

with exact elements  $a_{ij}$  and  $b_i$ , by one of the direct methods described in section 8.4. It may therefore appear to be pointless to consider any other method, such as an iterative method in which we make a guess at the solution and then refine it. However, iterative methods are important; we discuss them for the following reasons. First, it often happens that, when there is a large system of equations to be solved, a large number of the elements of the matrix of coefficients are zero; for example,

$$\begin{pmatrix} \times & \times & 0 & 0 & \times & 0 \\ \times & \times & 0 & \times & 0 & 0 \\ 0 & \times & \times & 0 & 0 & 0 \\ \times & 0 & 0 & \times & 0 & \times \\ \times & 0 & 0 & 0 & \times & 0 \\ 0 & 0 & \times & 0 & 0 & \times \end{pmatrix},$$

where the  $\times$ 's represent non-zero numbers. Such matrices are called **sparse matrices** (for obvious reasons). An iterative method, by taking account of the zeros from the outset, can sometimes give the result much more quickly than a direct method. Of course, if the pattern of zeros were convenient, for example, if we started with

$$\begin{pmatrix} \text{ } & \text{ } & \text{ } \\ \text{ } & \text{ } & \text{ } \\ \text{ } & \text{ } & \text{ } \end{pmatrix}$$

(A diagram showing a square matrix with a diagonal line from the top-left to the bottom-right. The upper triangular part (above the diagonal) is labeled 'x's' and the lower triangular part (below the diagonal) is labeled '0's'. The entire matrix is enclosed in large parentheses.)

then the back-substitution in the Gauss elimination method would be available straight away. The point is that the iterative method does not depend on a convenient pattern. Secondly, the iterative method, if it converges, improves the accuracy of the approximate solution at each



stage. There may be other advantages in terms of time and economy in the use of space in a digital computer, but these would need closer analysis.

In terms of automatic computation, iterative methods often have the advantage that various stages in the iteration repeat the *same* relatively simple process, which is ideal for economic and efficient programming. The iterative method is also of interest from a purely mathematical viewpoint; it shows how the ideas of the mapping of numerical error intervals, introduced in Volume 2, Chapter 7, can be extended to the discussion of vectors.

### Some Methods of Iteration

We shall use an illustration of two simultaneous equations in two unknowns although, customarily, the method would be used only on much larger systems.

We shall examine the system of equations

$$x_1 + 4x_2 = 6$$

$$2x_1 + x_2 = 5$$

The actual solution can easily be found to be

$$x_1 = 2, \quad x_2 = 1;$$

that is, the solution vector is

$$\begin{pmatrix} 2 \\ 1 \end{pmatrix}.$$

We postpone the complete matrix approach for a moment and simply rearrange the equations in a couple of ways. Firstly, we solve the first equation for  $x_2$  in terms of  $x_1$  and the second for  $x_1$  in terms of  $x_2$ , obtaining

$$x_2 = 1.5 - 0.25x_1$$

$$x_1 = 2.5 - 0.5x_2$$

We now develop the sequences  $x_1^{(r)}$ ,  $x_2^{(r)}$  where successive elements<sup>†</sup> are defined by

<sup>†</sup> To avoid confusion with the subscripts already in use and with the normal use of a superscript as an index, we use a superscript in brackets, e.g.  $x^{(r)}$ .



$$x_2^{(r+1)} = 1.5 - 0.25x_1^{(r)}$$

$$x_1^{(r+1)} = 2.5 - 0.5x_2^{(r)}$$

Let the first guess at the solution vector,

$$\begin{pmatrix} x_1^{(0)} \\ x_2^{(0)} \end{pmatrix},$$

be

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

We then get

$$x_2^{(1)} = 1.5 - 0.25 \times 0 = 1.5$$

$$x_1^{(1)} = 2.5 - 0.5 \times 0 = 2.5$$

and the sequence of estimated solution vectors up to the fifth term is

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2.5 \\ 1.5 \end{pmatrix}, \begin{pmatrix} 1.75 \\ 0.88 \end{pmatrix}, \begin{pmatrix} 2.06 \\ 1.06 \end{pmatrix}, \begin{pmatrix} 1.97 \\ 0.98 \end{pmatrix}.$$

Intuitively, it seems that this sequence is converging to the solution

vector  $\begin{pmatrix} 2 \\ 1 \end{pmatrix}$ .

### Exercise 1

Use the rearrangement

$$x_1 = 6 - 4x_2$$

$$x_2 = 5 - 2x_1,$$

starting with the vector  $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ , to obtain a sequence of five estimated solution

vectors for the system of equations in the text. Does it appear to converge?

The last exercise indicates that some rearrangements “work” and some do not. What we would like to have is some means of testing a rearrangement. For example, we might like to test



$$\begin{aligned}x_1 &= 3 + 0.5x_1 - 2x_2 \\x_2 &= 2.5 - x_1 + 0.5x_2,\end{aligned}$$

which is more complicated than either of the previous two rearrangements. We shall start our search for a “test” by examining the iterative process in more general terms. All our rearrangements have the iterative form

$$\underline{x}^{(r+1)} = G\underline{x}^{(r)} + H\underline{b}$$

where  $G$  and  $H$  are square matrices. For instance, in the rearrangement for which we calculated the sequence which appeared to converge, we had

$$\begin{aligned}x_1^{(r+1)} &= 2.5 - 0.5x_2^{(r)} \\x_2^{(r+1)} &= 1.5 - 0.25x_1^{(r)}\end{aligned}$$

This can be written in matrix form as

$$\begin{pmatrix} x_1^{(r+1)} \\ x_2^{(r+1)} \end{pmatrix} = \begin{pmatrix} 0 & -0.5 \\ -0.25 & 0 \end{pmatrix} \begin{pmatrix} x_1^{(r)} \\ x_2^{(r)} \end{pmatrix} + \begin{pmatrix} 0 & 0.5 \\ 0.25 & 0 \end{pmatrix} \begin{pmatrix} 6 \\ 5 \end{pmatrix}.$$

So in this case,

$$G = \begin{pmatrix} 0 & -0.5 \\ -0.25 & 0 \end{pmatrix} \quad \text{and} \quad H = \begin{pmatrix} 0 & 0.5 \\ 0.25 & 0 \end{pmatrix}.$$

We shall now attempt to explain what we mean by convergence of the sequence of estimated solution vectors. We define the **error vector** for the  $r$ th iteration to be

$$\underline{e}^{(r)} = \underline{x}^{(r)} - \underline{X},$$

where  $\underline{x}^{(r)}$  is the  $r$ th approximation to the exact solution vector  $\underline{X}$ . When the sequence converges to  $\underline{X}$ , the error vector approaches the zero vector. We require the error vector  $\underline{e}^{(r)}$  to get “smaller”, in some sense yet to be defined, as  $r$  increases.

The solution vector  $\underline{X}$  must satisfy the equation

$$\underline{X} = G\underline{X} + H\underline{b},$$

since this is merely a rearrangement of the equation

$$A\underline{X} - \underline{b} = \underline{0}.$$

From the two equations

$$\begin{aligned}\underline{x}^{(r+1)} &= G\underline{x}^{(r)} + H\underline{b} \\ \underline{X} &= G\underline{X} + H\underline{b},\end{aligned}$$



we can obtain a relationship between successive error vectors. We have

$$\begin{aligned}\underline{x}^{(r+1)} - \underline{X} &= G\underline{x}^{(r)} - G\underline{X} \\ &= G(\underline{x}^{(r)} - \underline{X}),\end{aligned}$$

i.e.

$$\underline{e}^{(r+1)} = G\underline{e}^{(r)}.$$

At first glance, it may seem that it is easy enough to judge whether  $\underline{e}^{(r+1)}$  is “smaller” than  $\underline{e}^{(r)}$ ; for instance, we can say that the individual elements of  $\underline{e}^{(r+1)}$  must all be smaller in absolute magnitude than the corresponding elements of  $\underline{e}^{(r)}$ . But, on reflection, this is seen to be unsatisfactory. What if *nearly all* the elements of  $\underline{e}^{(r+1)}$  are smaller than the corresponding elements of  $\underline{e}^{(r)}$ , but a few are just a little bigger? Looking at the individual elements will not do: we must find a *single number* as a measure.

### Exercise 2

Two other rearrangements of our original equations

$$x_1 + 4x_2 = 6$$

$$2x_1 + x_2 = 5$$

are

$$(i) \quad x_1 = 6 - 4x_2$$

$$x_2 = 5 - 2x_1$$

$$(ii) \quad x_1 = 3 + 0.5x_1 - 2x_2$$

$$x_2 = 2.5 - x_1 + 0.5x_2$$

Write down the corresponding iteration formulas in the matrix form

$$\underline{x}^{(r+1)} = G\underline{x}^{(r)} + H\underline{b},$$

and calculate the first five error vectors in each case from the equation

$$\underline{e}^{(r+1)} = G\underline{e}^{(r)},$$

starting with  $\underline{e}^{(0)} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$ .

### “Measuring” a Vector

To study the convergence of the iterative method we must have some measure by which we can test whether the error vector is getting “smaller”



at successive iterations. We can invent a measure in any way we like, and convergence may well depend on the measure we choose. So we must give some thought to what sort of measure is reasonable.

In the first place, we want our measure to be a *single number* because it is easy to compare numbers and to decide whether the sequence of such numbers, obtained from successive error vectors, is convergent to the number associated with the zero error vector. This means that, in the general case, we want to define a function which maps  $R^n$  to  $R$ .

We begin by considering some unsatisfactory measures, so that we can get a clearer idea of the conditions which a suitable “measure” must satisfy.

$$\text{Let } \underline{a} = \begin{pmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ \cdot \\ a_n \end{pmatrix} \text{ and } 0 = \begin{pmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{pmatrix}, \text{ the zero error vector.}$$

Let us define the function

$$\underline{a} \longmapsto a_1 \quad (\underline{a} \in R^n),$$

so that  $a_1$  is a “measure” of  $\underline{a}$ .

This measure is clearly unsatisfactory, for, although the sequence of  $a_1$ 's obtained from successive error vectors may converge to zero, this fact does not tell us what is happening to the remaining elements in the error vectors.

Consider the function

$$\underline{a} \longmapsto \sum_{i=1}^n a_i = a_1 + a_2 + \dots + a_n \quad (\underline{a} \in R^n).$$

Then,

$$\underline{0} \longmapsto 0,$$

and we want to see if the sequence of “measures” obtained from successive error vectors converges to zero.



But again this is unsatisfactory, because, for instance, if  $n$  is even and the error vectors are

$$\underline{e}^{(r)} = \begin{pmatrix} 1 \\ -1 \\ 1 \\ \vdots \\ -1 \end{pmatrix}$$

for all  $r > N$ , where  $N$  is some positive integer, then  $\underline{e}^{(r)} \rightarrow 0$ , but the error in each element of the estimated solution to our system is not negligible.

The first measure is unsatisfactory because it does not take account of all the elements in the error vector. The second measure is unsatisfactory because, although it does take account of all the elements of the vector, we can get a misleading result if some of the elements are positive and some are negative.

We can improve on the second measure in two fairly obvious ways: we define the functions

$$m_1: \underline{a} \mapsto \left( \sum_{i=1}^n a_i^2 \right)^{1/2} = (a_1^2 + \cdots + a_n^2)^{1/2} \quad (\underline{a} \in R^n)$$

and

$$m_2: \underline{a} \mapsto \sum_{i=1}^n |a_i| = |\underline{a}_1| + \cdots + |\underline{a}_n| \quad (\underline{a} \in R^n).$$

The measure defined by  $m_1$  is the more obvious for two reasons: the measure is the same as the standard deviation (used in statistics) of the  $a_i$  from zero; also, in two or three dimensions, this measure can be interpreted as the length of the corresponding geometric vector.

We shall look at each of these measures briefly to illustrate their use, restricting ourselves to the geometric interpretation in two or three dimensions.

We look again at the two-dimensional example we discussed above and, in particular, at the rearrangement which led to the equation

$$\underline{e}^{(r+1)} = G\underline{e}^{(r)}$$

where

$$G = \begin{pmatrix} 0 & -0.5 \\ -0.25 & 0 \end{pmatrix}.$$



If we write

$$\underline{e}^{(r)} = \begin{pmatrix} e_1^{(r)} \\ e_2^{(r)} \end{pmatrix},$$

then

$$m_1: \underline{e}^{(r)} \longmapsto \sqrt{(e_1^{(r)})^2 + (e_2^{(r)})^2}.$$

If we represent the error vector by an arrow from the origin of a set of Cartesian co-ordinates to the point with co-ordinates  $(e_1^{(r)}, e_2^{(r)})$ , then  $m_1(e^{(r)})$  is the length of the arrow. Now, we have

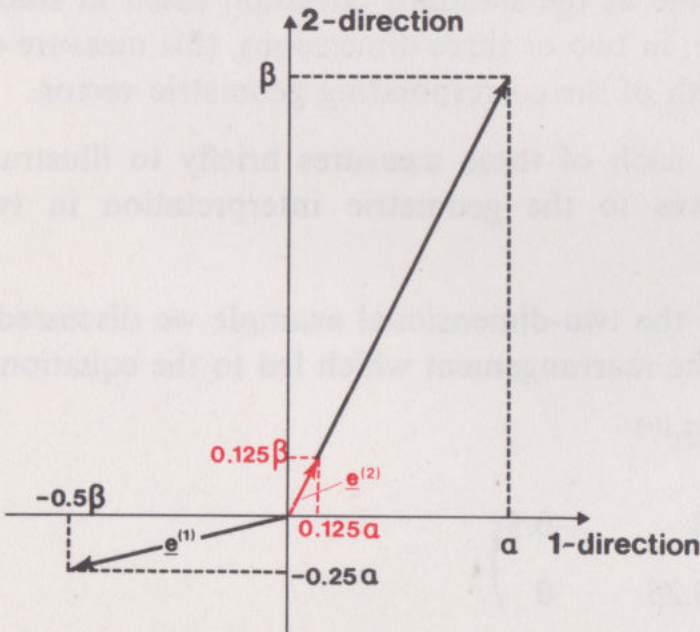
$$\begin{pmatrix} e_1^{(r+1)} \\ e_2^{(r+1)} \end{pmatrix} = \begin{pmatrix} 0 & -0.5 \\ -0.25 & 0 \end{pmatrix} \begin{pmatrix} e_1^{(r)} \\ e_2^{(r)} \end{pmatrix},$$

that is,

$$\begin{aligned} e_1^{(r+1)} &= -0.5 e_2^{(r)} \\ e_2^{(r+1)} &= -0.25 e_1^{(r)}. \end{aligned}$$

Let us suppose that the initial error vector  $\underline{e}^{(0)} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$ ; then we can draw an arrow from the origin to represent the error vector. The successive error vectors are then represented as follows.

$r$	$e_1^{(r)}$	$e_2^{(r)}$
0	$\alpha$	$\beta$
1	$-0.5\beta$	$-0.25\alpha$
2	$0.125\alpha$	$0.125\beta$





In this particular example, the error vector has one of two directions and alternates between them, so that, after two steps in the iteration, the error vector is again pointing in the same direction, but its length has been decreased by a factor of 8. By taking a sufficient number of steps we can make the length of the error vector as small as we please, i.e. we can get the pointed end of the arrow as close to the origin as we please. This suggests that, with this measure of an error vector, the iteration method will converge to a solution whatever error  $\begin{pmatrix} \alpha \\ \beta \end{pmatrix}$  there is in our initial guess which starts the iteration.

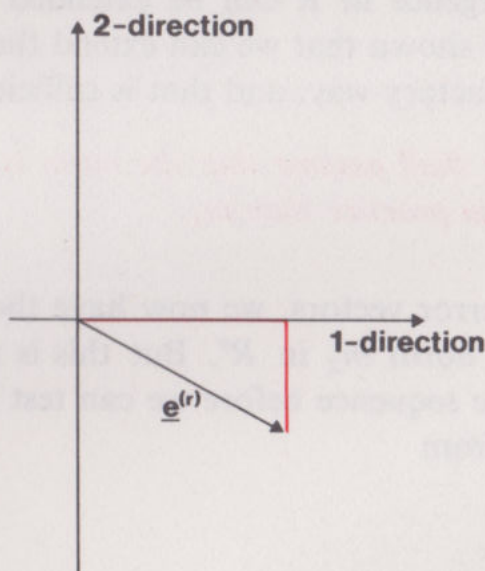
We now consider the geometrical interpretation of the second measure, defined by:

$$m_2(\underline{a}) = \sum_{i=1}^n |a_i|.$$

With the notation as before, in two dimensions we have

$$m_2(\underline{e}^{(r)}) = |e_1^{(r)}| + |e_2^{(r)}|,$$

which is represented as the sum of the two *lengths* marked in red on the diagram.



In our example, we see that the sum of the moduli for successive error vectors gets smaller and smaller, so that once again we have an intuitive concept of the error vector converging to the zero vector.

Having gained an intuitive impression of what we mean by convergence for vectors, we shall now be more precise.



A measure of the kind we have been discussing is called a *norm*. A **norm** is a function with certain special properties which maps the elements of a vector space  $V$  to  $R$ . The image of a vector  $\underline{v}$  under this function is denoted by  $||\underline{v}||$  (read as “the norm of  $\underline{v}$ ”).

A norm is defined to have the following properties:

- (i)  $||\underline{v}|| \geq 0$ , for all  $\underline{v} \in V$ ;
- (ii)  $||\underline{v}|| = 0$  if and only if  $\underline{v} = \underline{0}$ ;
- (iii)  $||\underline{v}_1 + \underline{v}_2|| \leq ||\underline{v}_1|| + ||\underline{v}_2||$ , for all  $\underline{v}_1, \underline{v}_2 \in V$ ;
- (iv)  $||\alpha \underline{v}_1|| = |\alpha| ||\underline{v}_1||$ , where  $\alpha$  is any real number.

We now define convergence in a vector space as follows. Let  $m$  be a norm on the vector space, and let

$$\underline{x}^{(1)}, \underline{x}^{(2)}, \dots$$

be an infinite sequence of vectors. This sequence is said to have the limit  $\underline{X}$  if the sequence of norms of the error vectors:

$$||\underline{x}^{(1)} - \underline{X}||, ||\underline{x}^{(2)} - \underline{X}||, \dots$$

has limit 0. So we have defined convergence in a vector space in terms of the more familiar notion of convergence in  $R$ .† It is not our intention to study norms further, but an obvious next step would be to see which of our results for convergence in  $R$  can be extended to convergence in a vector space. We have shown that we can extend the idea of convergence in an apparently satisfactory way, and that is sufficient for now.

*In subsequent work we shall assume that the norm is  $m_2$ , since this proves to be more convenient in practice than  $m_1$ .*

Given a sequence of error vectors, we now have the means to test it for convergence using the norm  $m_2$  in  $R^n$ . But this is not ideal, because we first have to obtain the sequence before we can test it. We know that the sequence is obtained from

$$\underline{e}^{(r+1)} = G\underline{e}^{(r)}$$

if our original equation

$$A\underline{x} = \underline{b}$$

was arranged in the form

$$\underline{x}^{(r+1)} = G\underline{x}^{(r)} + H\underline{b}.$$

† Convergence in  $R$  was discussed in Volume I.



In other words, the sequence of error vectors is determined by the rearrangement; in particular, it is determined by the matrix  $G$ . So we must now determine what conditions we can impose on  $G$  (i.e. on the rearrangement) so that the resulting sequence of error vectors converges to  $\underline{0}$ .

Suppose now that, for all  $r$ ,

$$m_2(\underline{e}^{(r+1)}) < m_2(\underline{e}^{(r)}), \quad \text{Inequality (1)}$$

i.e.

$$|e_1^{(r+1)}| + |e_2^{(r+1)}| + \dots + |e_n^{(r+1)}| < |e_1^{(r)}| + |e_2^{(r)}| + \dots + |e_n^{(r)}|.$$

This inequality is true for all  $r$ , so we can find a number  $k$ ,  $0 < k < 1$ , such that

$$\begin{aligned} \sum_{i=1}^n |e_i^{(r+1)}| &\leq k \sum_{i=1}^n |e_i^{(r)}| \\ &\leq k^2 \sum_{i=1}^n |e_i^{(r-1)}|. \\ &\dots \\ &\leq k^r \sum_{i=1}^n |e_i^{(1)}|. \end{aligned}$$

It follows that

$$\lim_{r \text{ large}} \left( \sum_{i=1}^n |e_i^{(r+1)}| \right) \leq \left( \lim_{r \text{ large}} k^r \right) \times \left( \sum_{i=1}^n |e_i^{(1)}| \right),$$

since  $\sum_{i=1}^n |e_i^{(1)}|$  is a constant.

Now, since  $0 < k < 1$ ,  $\lim_{r \text{ large}} k^r = 0$ , so that

$$\lim_{r \text{ large}} \left( \sum_{i=1}^n |e_i^{(r+1)}| \right) = 0.$$

So we see that if Inequality (1) is satisfied the sequence of error vectors converge. Inequality (1) can be written in the alternative form

$$\frac{m_2(\underline{e}^{(r+1)})}{m_2(\underline{e}^{(r)})} < 1.$$

We know that

$$\underline{e}^{(r+1)} = G\underline{e}^{(r)},$$



so that this condition becomes

$$\frac{m_2(G\bar{e}^{(r)})}{m_2(\bar{e}^{(r)})} < 1,$$

and it looks as if the clue to convergence is held by  $G$ . The information contained in  $G$  will relate the norms of successive error vectors in some way.

Suppose

$$G = \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix}.$$

Successive error vectors are then

$$\begin{pmatrix} e_1^{(r+1)} \\ e_2^{(r+1)} \end{pmatrix} = \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix} \begin{pmatrix} e_1^{(r)} \\ e_2^{(r)} \end{pmatrix},$$

so

$$e_1^{(r+1)} = g_{11}e_1^{(r)} + g_{12}e_2^{(r)}$$

and

$$e_2^{(r+1)} = g_{21}e_1^{(r)} + g_{22}e_2^{(r)}.$$

Thus

$$\begin{aligned} m_2(\bar{e}^{(r+1)}) &= |e_1^{(r+1)}| + |e_2^{(r+1)}| \\ &= |g_{11}e_1^{(r)} + g_{12}e_2^{(r)}| + |g_{21}e_1^{(r)} + g_{22}e_2^{(r)}| \\ &\leq |g_{11}e_1^{(r)}| + |g_{12}e_2^{(r)}| + |g_{21}e_1^{(r)}| + |g_{22}e_2^{(r)}|, \end{aligned}$$

using the triangle inequality (see section 4.3).

Since, for any real numbers  $\alpha, \beta$ ,

$$|\alpha \times \beta| = |\alpha| \times |\beta|,$$

we have

$$|e_1^{(r+1)}| + |e_2^{(r+1)}| \leq |g_{11}||e_1^{(r)}| + |g_{12}||e_2^{(r)}| + |g_{21}||e_1^{(r)}| + |g_{22}||e_2^{(r)}|,$$

and finally

$$|e_1^{(r+1)}| + |e_2^{(r+1)}| \leq (|g_{11}| + |g_{21}||e_1^{(r)}| + (|g_{12}| + |g_{22}||e_2^{(r)}|).$$

Therefore if we can find a number  $k$  in the range  $0 < k < 1$  such that both

$$|g_{11}| + |g_{21}| \leq k$$



and

$$|g_{12}| + |g_{22}| \leq k,$$

the convergence condition is satisfied, and the sequence of error vectors converges to the zero vector. Thus we can test the matrix  $G$  by adding the moduli of the elements in each column. If the largest column sum is less than 1, the sequence of error vectors converges. For example, if

$$G = \begin{pmatrix} 0.3 & 0.5 \\ 0.2 & 0.1 \end{pmatrix},$$

then we have

$$|g_{11}| + |g_{21}| = 0.5 \quad \text{and} \quad |g_{12}| + |g_{22}| = 0.6.$$

If we take  $k = 0.6$  (or any number between 0.6 and 1), then the convergence condition is satisfied, so the iterative sequence produced by the matrix  $G$  converges.

(This number  $k$ , representing the maximum sum of the moduli of the elements of a column of a matrix, can itself be regarded as a norm of a matrix. In other words, it is a method of “measuring” or attaching a number to a matrix.)

### Exercise 3

Test each of the following matrices to see if the sequence of error vectors obtained from the equation

$$\underline{e}^{(r+1)} = G\underline{e}^{(r)}$$

is certain to converge to the zero vector. If it is, give a suitable value for the constant  $k$ .

(i)  $G = \begin{pmatrix} 0.3 & 0.75 \\ 0.6 & 0.2 \end{pmatrix}$

(ii)  $G = \begin{pmatrix} 0.7 & 0.05 \\ -0.5 & 0.1 \end{pmatrix}$

(iii)  $G = \begin{pmatrix} 0.4 & 0.3 \\ 0.6 & 0.7 \end{pmatrix}$

(iv)  $G = \begin{pmatrix} 0.33 & -0.72 \\ -0.66 & -0.27 \end{pmatrix}$



*Exercise 4*

Rearrange the following equations in such a way that the resulting iterative method converges. Hence solve the equations by iteration.

$$5x_1 + 3x_2 = 7$$

$$2x_1 - 4x_2 = 3.$$

The convergence test can be applied to a sequence of error vectors in a vector space of any finite dimension. We sum the moduli of the elements of each column of a matrix, and if all the sums are less than or equal to  $k$ , which is a positive number  $< 1$ , then we have guaranteed convergence.

Finally, we turn to a point made in the introduction to this section. We said that the iterative method is useful when the matrix of coefficients is a sparse matrix. This is so because when there are only a few non-zero elements in each row, we can readily isolate one of the elements of the solution vector and express it in terms of relatively few elements on the right-hand side. (There may be some manipulation required to obtain such an equation for *each* element.) For example, the arrangement

$$x_1 = 0.2x_2 + 0.5x_5 + 1$$

$$x_2 = 0.4x_1 + 0.1x_3 + 2$$

$$x_3 = 0.2x_2 + 0.3x_4 + 3$$

$$x_4 = 0.5x_3 + 0.2x_5 + 4$$

$$x_5 = 0.4x_1 + 0.3x_2 + 5$$

has 3 fewer multiplications on the right-hand side of each equation than it might have. This is because the matrix of coefficients written in the form

$$\begin{pmatrix} 1 & -0.2 & 0 & 0 & -0.5 \\ -0.4 & 1 & -0.1 & 0 & 0 \\ 0 & -0.2 & 1 & -0.3 & 0 \\ 0 & 0 & -0.5 & 1 & -0.2 \\ -0.4 & -0.3 & 0 & 0 & 1 \end{pmatrix}$$

is relatively sparse. The fact that the number of multiplications required is drastically reduced would be even more pronounced if, for example, there were only 5 non-zero elements in each row of a  $100 \times 100$  matrix. The final judgment on efficiency, compared with a direct method, is



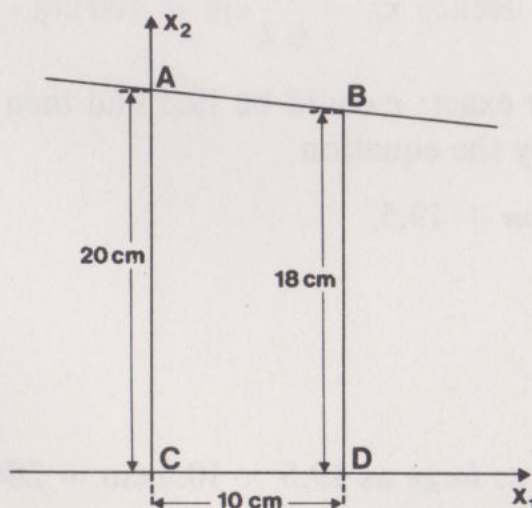
more difficult however, for we also have to decide how fast the iterative process converges, assuming it does converge; that is how many steps are required to obtain the required accuracy.

## 8.6 Ill-conditioned Systems of Equations

In this section we look at a particularly disastrous way in which errors can sometimes accumulate when we solve simultaneous equations based on inexact data or when we introduce rounding errors during solution. In fact, the error accumulation may even render the results of solving the simultaneous equations virtually useless. When small changes in the data have a large effect on the solution of a system of equations, then we say that the system is **ill-conditioned**. This term has no precise definition but is used in the same relative sense that adjectives like “small” are used in English. “Small changes in the data” producing “large changes in the result” is simply one way of describing the phenomenon of ill-conditioning; the adjectives used indicate the imprecision of the concept. A numerical illustration of ill-conditioning is given in the following example.

### Example 1

In a reconstruction of a crime, bullet holes centred at  $A$  and  $B$  were found in a double-glazed window at distances apart indicated on the diagram, the measurements being taken to the centres of the holes. All the measurements were taken *to the nearest centimetre*. Other evidence showed that the gun was at the level  $CD$  when fired. How accurately can we determine the position from which the gun was fired?

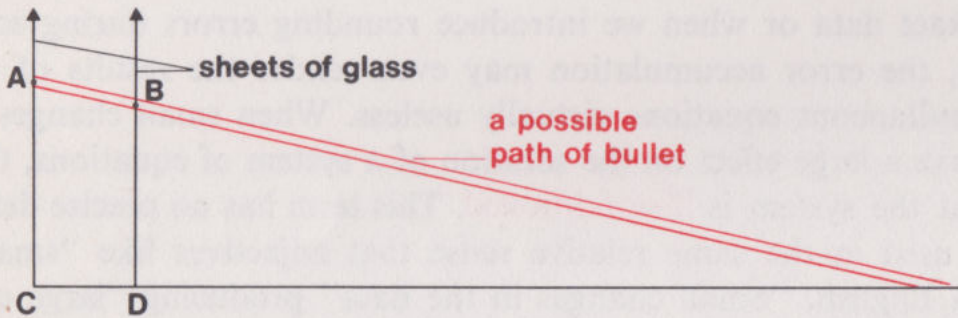




Before following through the solution to this example, you may find it helpful to assume that the measurement of 10 cm is *exact* and draw accurately, on a piece of graph paper, the two most extreme possible trajectories of the bullet, assuming them to be straight lines.

### *Solution of Example 1*

If we fit  $x_1$  and  $x_2$  co-ordinate axes on to the diagram, we are effectively finding the intersection of the straight line  $AB$  with  $x_1$ -axis  $CD$ .



The equation of the straight line  $AB$  is

$$x_2 = mx_1 + c,$$

where (in theory) we can find  $m$  and  $c$  from the data. The distance  $x_G$  from  $C$  at which the bullet was fired is the value of  $x_1$ , when  $x_2 = 0$ , i.e.

$$x_G = -\frac{c}{m}.$$

If we assume the data to be exact, then we know that  $AB$  passes through the points  $(0, 20)$  and  $(10, 18)$ , so that

$$20 = c$$

$$18 = 10m + c,$$

whence  $m = -0.2$ . Hence  $x_G = \frac{20}{0.2} \text{ cm} = 100 \text{ cm} = 1 \text{ m}$ .

But the data are *not* exact:  $c$  could be 19.5 and then  $m$  could be as *small* as the value given by the equation

$$18.5 = 10.5m + 19.5,$$

i.e.

$$m = \frac{-1}{10.5};$$

so that  $x_G$  could be as *large* as  $19.5 \times 10.5 \text{ cm} = 204.75 \text{ cm}$ . That is, the

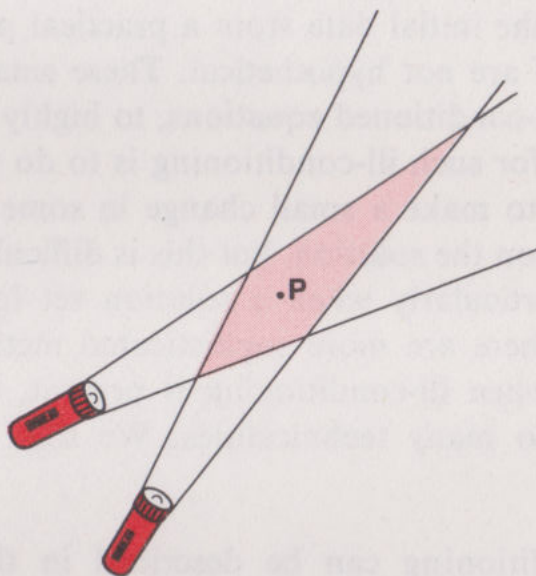


approximate value for  $x_G$  (1 m) could be over 1 m in error. In fact, all we can say is that

$$x_G \in [64.92, 204.75] \text{ cm.}$$

In other words, the distance the gun was fired from the window could be anything between just over 0.5 m and just over 2 m. This is not a very accurate result when you consider the apparent accuracy of the original measurements.

In geometric terms, in the last example we were trying to find the intersection of a pair of nearly parallel straight lines (the  $x_1$ -axis was one of these). If there is a small error in the original data, it can result in a considerable error in the solution. One possible way of envisaging what is happening is to imagine two torches, or searchlights, with their beams crossed.



The exact solution, if the data are exact, is represented by a point  $P$ . With inaccuracies in the data, represented by the diverging beams from the torches, the straight lines could be anywhere within the beams. The point representing the true solution could be anywhere within the “area of intersection” which is shaded in the diagram.

Intuitively, the more nearly parallel to each other the axes of the torch beams become, the larger the shaded area, and the further away the true solution can be from the approximate solution.

Now let us look at a more general case in two dimensions in the following exercise.



*Exercise 1*

Solve the following simultaneous equations for  $x_1$  and  $x_2$ :

$$x_1 + x_2 = 1$$

$$x_1 + (1 + \epsilon)x_2 = 2,$$

where  $\epsilon$  is some real number, by the Gauss elimination method. Write down the solutions corresponding to

(i)  $\epsilon = 0.01, 0.02$

(ii)  $\epsilon = 2.01, 2.04$

respectively. Compare the percentage changes in the *coefficient* of  $x_2$  in the second equation for cases (i) and (ii) with the percentage changes in the corresponding *solution* for  $x_2$ .

This last exercise was rather hypothetical. However, the rounding errors which occur both in the initial data from a practical problem and in the solution process itself are not hypothetical. These small rounding errors can readily lead, in ill-conditioned equations, to highly inaccurate results. A simple way to test for such ill-conditioning is to do what we did in the last exercise, that is, to make a small change in some coefficients to see what effects there are on the solution, but this is difficult to do with larger sets of equations, particularly when a solution set looks acceptable in the physical sense. There are more sophisticated methods which give a practical indication when ill-conditioning is present, but most of these will involve us in too many technicalities. We shall describe just one such method.

*Theoretically*, ill-conditioning can be described in the following way. Given a system of simultaneous equations in matrix form

$$A\mathbf{x} = \mathbf{b},$$

the system is ill-conditioned if the  $n$  columns of the matrix are *almost* linearly dependent (note the vagueness again), or, in other words, if the matrix is *nearly singular*. This means that all the elements of  $A$  are close to the corresponding elements of some singular matrix. In the last exercise, for example,  $A$  would be

$$\begin{pmatrix} 1 & 1 \\ 1 & 1 + \epsilon \end{pmatrix}.$$

Clearly if  $\epsilon$  were small (we found this made the equations ill-conditioned),



a small change, that is a change of  $-\epsilon$  in  $a_{22}$ , would turn the matrix into the singular matrix

$$\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

in which the two columns are obviously linearly dependent.

### Exercise 2

Determine the inverse matrix  $A^{-1}$  of

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 + \epsilon \end{pmatrix}.$$

What are the elements of  $A^{-1}$  if  $\epsilon = 0.01$ ?

This last exercise shows one feature of the inverse of the matrix of coefficients of an ill-conditioned system of equations; that is, its elements are relatively large when compared with the elements of the original matrix. This leads to one final point about systems of equations which may be ill-conditioned. A recommended and useful means of checking the accuracy of an estimated solution  $\underline{x}_s$  of

$$A\underline{x} = \underline{b}$$

is to premultiply  $\underline{x}_s$  by  $A$  and obtain a vector  $\underline{b}_s$ . That is,

$$A\underline{x}_s = \underline{b}_s.$$

If the elements of  $\underline{b}_s$  differ very little from the elements of  $\underline{b}$ , then we would normally assume that the solution is fairly accurate. (Remember that there will almost always be some inaccuracies from the rounding errors in a real computation of any length.) If the system of equations is ill-conditioned, however, the solutions can still be grossly in error; for if we subtract the two equations above we get

$$A(\underline{x}_s - \underline{x}) = \underline{b}_s - \underline{b},$$

or, using the error vector notation,

$$A\underline{e} = \underline{E}$$

where

$$\underline{e} = \underline{x}_s - \underline{x} \quad \text{and} \quad \underline{E} = \underline{b}_s - \underline{b}.$$

Then we shall have

$$\underline{e} = A^{-1}\underline{E}.$$



If  $A^{-1}$  contains large elements, it is intuitively obvious that  $\underline{e}$  can be large even though  $\underline{E}$  is small. So, whenever we suspect ill-conditioning, it is worth having a look at the size of the elements of  $A^{-1}$  relative to the elements of  $A$ .

### Exercise 3

If

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1.01 \end{pmatrix},$$

find an  $\underline{E}$  with norm (sum of the moduli of the elements)  $< 0.01$  which leads to an  $\underline{e}$  with norm  $> 1$ .

Finally, there is the problem of what to do about ill-conditioning when it occurs and we do recognize it. It may well be that reformulation of the problem in terms of different variables, as may be possible with sets of ill-conditioned simultaneous equations arising from problems involving engineering structures, will lead to a well-conditioned set of equations. Otherwise we can use double precision arithmetic in the calculation, that is, carry twice as many digits throughout the work, in an attempt to get better results, but if the equations are badly ill-conditioned this will still be of no avail. Then the advice is—forget the problem, or quote the result to whatever accuracy is obtained, however bad.

### Summary

When the matrix of coefficients is nearly singular, that is, a small change in some elements of the matrix will produce a singular matrix, the matrix equation

$$A\underline{x} = \underline{b}$$

is said to be ill-conditioned. This usually means that the solution is highly unreliable, particularly if the original data, from which the matrix equation arose, were inexact.

This chapter as a whole should be regarded as a very brief introduction to numerical methods in this area of mathematics. There are many refinements and developments which would modify some of our conclusions.



## 8.7 Additional Exercises

### Exercise 1

If  $A$  and  $B$  are any  $n \times n$  matrices, prove that

$$r(BA) \leq r(A).$$

### Exercise 2

The simultaneous equations

$$x_1 + 2x_2 = 4$$

Equations (1)

$$1000x_1 + 2001x_2 = 4003$$

have solution vector  $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -2 \\ 3 \end{pmatrix}$

(i) Solve the following two systems of simultaneous equations

$$x_1 + 2x_2 = 4$$

Equations (2)

$$1000x_1 + 1999x_2 = 4003$$

$$x_1 + 2x_2 = 4$$

Equations (3)

$$1000x_1 + 2003x_2 = 4003$$

(ii) Hence discuss the validity of the solution of Equations (1).

### Exercise 3

The system of simultaneous equations

$$5x_1 + x_2 + x_3 = 1$$

$$x_1 + 10x_2 + x_3 = 1$$

$$x_1 + x_2 + 5x_3 = 1$$

can be re-arranged as

$$x_1 = 0.2 - 0.2x_2 - 0.2x_3$$

$$x_2 = 0.1 - 0.1x_1 - 0.1x_3$$

$$x_3 = 0.2 - 0.2x_1 - 0.2x_2$$

(i) Explain why this rearrangement gives rise to an iterative process which will converge, by discussing the matrix involved in the iterative process.



(ii) Obtain the solution vector

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0.155 \\ 0.069 \\ 0.155 \end{pmatrix}$$

in which each element is correct to three decimal places by the iterative process, starting with the guess

$$\begin{pmatrix} 0.2 \\ 0.1 \\ 0.2 \end{pmatrix}$$

or any other guess that you prefer.

N.B. The *exact* solution is

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 9/58 \\ 2/29 \\ 9/58 \end{pmatrix}$$

#### Exercise 4

Solve by the Gauss elimination method the following system of simultaneous equations.

$$\begin{aligned} 2x_1 - 3x_2 - 4x_3 &= -1 \\ x_1 + 3x_2 + 2x_3 &= 2 \\ -3x_1 + 6x_2 - 2x_3 &= 0 \end{aligned}$$

#### Exercise 5

Find the inverse of the matrix

$$A = \begin{pmatrix} 2 & 1 & 1 \\ 3 & 2 & 1 \\ 4 & 2 & 1 \end{pmatrix}$$

by using elementary row operations. Incorporate a row sum check.



## 8.8 Answers to Exercises

### Section 8.1

#### Exercise 1

$$(i) \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

$$(ii) \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$(iii) \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & 3 & 1 \end{pmatrix}$$

$$(iv) \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix}$$

#### Exercise 2

The  $E$ 's are not unique: they will depend on which elementary operations are chosen. But  $P$  is unique:

$$\begin{pmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ -\frac{11}{4} & \frac{1}{4} & 1 \end{pmatrix}$$

A possible choice of elementary matrices is

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{1}{4} & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -2 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ -\frac{11}{4} & \frac{1}{4} & 0 \end{pmatrix}$$

$E_3 \qquad E_2 \qquad E_1 \qquad P$

### Section 8.2

#### Exercise 1

Just multiply out.



*Exercise 2*

$$(i) \begin{pmatrix} \frac{7}{8} & \frac{1}{16} & -\frac{1}{32} \\ -\frac{1}{8} & \frac{1}{16} & -\frac{1}{32} \\ -\frac{7}{4} & \frac{3}{8} & \frac{5}{16} \end{pmatrix}.$$

$$(ii) \begin{pmatrix} -\frac{30}{23} & \frac{5}{23} & -\frac{8}{23} \\ \frac{17}{23} & \frac{1}{23} & \frac{3}{23} \\ -\frac{6}{23} & \frac{1}{23} & \frac{3}{23} \end{pmatrix}.$$

**Section 8.3***Exercise 1*

The rank of the matrix is 3.

*Exercise 2*

Since  $B$  is non-singular, we can apply the same argument to  $B$  as to  $E$  in the text preceding Example 1. That is,

$$r(BA) = \text{dimension of } B(A(R_n)) = \text{dimension of } A(R^n) = r(A).$$

**Section 8.4***Exercise 1*

8



Exercise 2

Step in the Calculation	Number of Multiplications and/or Divisions
Eliminate $x_1$ from equations (2) and (3).	
Factor for equation (2) is $-0.6$	1 division
Factor for equation (3) is $-0.4$	1 division
$5x_1 + 2x_2 + (-2) \quad x_3 = 8 \quad (1)$	
$4.8 \quad x_2 + -2.8 \quad x_3 = -3.8 \quad (4)$	3 multiplications
$3.2 \quad x_2 + -1.2 \quad x_3 = -2.2 \quad (5)$	3 multiplications
Now eliminate $x_2$ from equation (5).	
Factor for equation (5) is $-0.66$	1 division
$5x_1 + 2x_2 + (-2) \quad x_3 = 8 \quad (1)$	
$4.8 \quad x_2 + -2.8 \quad x_3 = -3.8 \quad (4)$	
$0.66 \quad x_3 = 0.33 \quad (6)$	2 multiplications
Therefore	
$x_3 = 0.5$	1 division
and back-substituting into equation (4) gives	
$x_2 = -0.5$	1 multiplication and 1 division
and from equation (1)	
$x_1 = 2$	2 multiplications and 1 division
Total number of divisions and multiplications	17

Notice that the number of multiplications and divisions is roughly half the corresponding number for the method considered at the beginning of this section.



## Exercise 3

Step in the Calculation	Number of Multiplications/Divisions
Eliminate $x_1$ from all equations after the first.	$(n - 1)(n + 1) = n^2 - 1$
Eliminate $x_2$ from all equations after the second.	$(n - 2)n = (n - 1)^2 - 1$
.	
.	
Eliminate $x_{n-1}$ from all equations after the $(n - 1)$ th, i.e. the $n$ th equation.	$2^2 - 1$
Total for elimination is	$\left( \sum_{r=2}^n r^2 \right) - (n - 1)$ $= \left( \frac{n(n + 1)(2n + 1)}{6} - 1 \right) - (n - 1)$
Now carry out the back substitution.	
To determine $x_n$ from an equation such as $\alpha_n x_n = \beta_n$	1
To determine $x_{n-1}$ from an equation such as $\gamma_{n-1} x_{n-1} + \gamma_n x_n = \beta_{n-1}$	2
.	
.	
To determine $x_1$	$n$
Total for back substitution is	$\sum_{r=1}^n r = \frac{n(n + 1)}{2}$
Therefore total number of operations is	$\frac{n^3}{3} + n^2 - \frac{n}{3}$



## Exercise 4

$$\begin{pmatrix} \boxed{n \text{ terms}} \\ \vdots \\ \vdots \end{pmatrix} \begin{pmatrix} \boxed{n \text{ terms}} & \cdots & \cdots \end{pmatrix}$$

Each element of the product matrix requires  $n$  multiplications.

There are  $n^2$  terms in the product matrix. Therefore the number of multiplications is  $n^3$ .

## Section 8.5

## Exercise 1

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 6 \\ 5 \end{pmatrix}, \begin{pmatrix} -14 \\ -7 \end{pmatrix}, \begin{pmatrix} 34 \\ 33 \end{pmatrix}, \begin{pmatrix} -126 \\ -63 \end{pmatrix}$$

It does not appear to converge.

## Exercise 2

$$(i) \underline{x}^{(r+1)} = \begin{pmatrix} 0 & -4 \\ -2 & 0 \end{pmatrix} \underline{x}^{(r)} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \underline{b}$$

The first five error vectors are

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \begin{pmatrix} -4\beta \\ -2\alpha \end{pmatrix}, \begin{pmatrix} 8\alpha \\ 8\beta \end{pmatrix}, \begin{pmatrix} -32\beta \\ -16\alpha \end{pmatrix}, \begin{pmatrix} 64\alpha \\ 64\beta \end{pmatrix}.$$

The error vectors are getting bigger by any reasonable “measure”, which suggests non-convergence; this agrees with our conclusions in Exercise 1.

$$(ii) \underline{x}^{(r+1)} = \begin{pmatrix} 0.5 & -2 \\ -1 & 0.5 \end{pmatrix} \underline{x}^{(r)} + \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix} \underline{b}$$

The first five error vectors are

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \begin{pmatrix} 0.5\alpha - 2\beta \\ -\alpha + 0.5\beta \end{pmatrix}, \begin{pmatrix} 2.25\alpha - 2\beta \\ -\alpha + 2.25\beta \end{pmatrix}, \begin{pmatrix} 3.125\alpha - 5.5\beta \\ -2.75\alpha + 3.125\beta \end{pmatrix}, \\ \begin{pmatrix} 7.0625\alpha - 9\beta \\ -4.5\alpha + 7.0625\beta \end{pmatrix}.$$



This time it is not so easy to see what is going on. For instance, if we choose  $\alpha = \beta = 1$ , we get the following error vectors:

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} -1.5 \\ -0.5 \end{pmatrix}, \begin{pmatrix} 0.25 \\ 1.25 \end{pmatrix}, \begin{pmatrix} -2.375 \\ 0.375 \end{pmatrix}, \begin{pmatrix} -1.9375 \\ 2.5625 \end{pmatrix}.$$

Again, we have a sequence of vectors which does not look very hopeful.

### Exercise 3

- (i) The sequence converges. Any  $k$  such that  $0.95 \leq k < 1$  is suitable.
- (ii) The sequence does not necessarily converge.
- (iii) The sequence does not necessarily converge.
- (iv) The sequence converges. Any  $k$  such that  $0.99 \leq k < 1$  is suitable.

Note that we have only shown that

condition satisfied implies convergence

and not that

condition not satisfied implies divergence;

we may still have convergence in the latter case, although it may be more difficult to prove. Thus you might like to look at the sequence generated by the  $G$ 's in cases (ii) and (iii) starting with  $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ .

### Exercise 4

One such rearrangement is

$$x_1 = 1.4 - 0.6x_2$$

$$x_2 = -0.75 + 0.5x_1.$$

With this rearrangement,

$$G = \begin{pmatrix} 0 & -0.6 \\ 0.5 & 0 \end{pmatrix}$$

and the iteration formula takes the form

$$\underline{x}^{(r+1)} = \begin{pmatrix} 0 & -0.6 \\ 0.5 & 0 \end{pmatrix} \underline{x}^{(r)} + \begin{pmatrix} 1.4 \\ -0.75 \end{pmatrix}.$$



Starting with any  $\underline{x}^{(1)}$ , we shall eventually get as close as we please to the

actual solution  $\begin{pmatrix} \frac{37}{26} \\ -\frac{1}{26} \end{pmatrix} \simeq \begin{pmatrix} 1.42 \\ -0.04 \end{pmatrix}.$

Section 8.6

Exercise 1

$$\begin{aligned} x_1 + x_2 &= 1 \\ x_1 + (1 + \epsilon)x_2 &= 2 \end{aligned}$$

Carrying out the Gauss elimination method gives

$$\begin{aligned} x_1 + x_2 &= 1 \\ \epsilon x_2 &= 1, \end{aligned}$$

which gives

$$x_2 = \frac{1}{\epsilon} \quad (\epsilon \neq 0),$$

and from back-substitution,

$$x_1 = 1 - \frac{1}{\epsilon}$$

	(i)		(ii)	
$\epsilon$	0.01	0.02	2.01	2.04
Solution to 2 decimal places	(−99, 100)	(−49, 50)	(0.50, 0.50)	(0.51, 0.49)

Percentage change of coefficient in  $x_2$  is 1% to one place of decimals in each case.

In (i) the solution  $x_2$  changes by 50%; in (ii) by 2%. This indicates that “when  $\epsilon$  is small the equations are ill-conditioned” means in this case “small compared with unity”.



*Exercise 2*

$$A^{-1} = \begin{pmatrix} 1 + \frac{1}{\epsilon} & -\frac{1}{\epsilon} \\ -\frac{1}{\epsilon} & \frac{1}{\epsilon} \end{pmatrix}$$

$$\begin{pmatrix} 101 & -100 \\ -100 & 100 \end{pmatrix}$$

*Exercise 3*

One example is

$$\underline{E} = \begin{pmatrix} -0.004 \\ 0.004 \end{pmatrix}$$

Then  $m_2(\underline{E}) = 0.008 < 0.01$  and

$$\underline{e} = \begin{pmatrix} 101 & -100 \\ -100 & 100 \end{pmatrix} \begin{pmatrix} -0.004 \\ 0.004 \end{pmatrix} = \begin{pmatrix} -0.804 \\ 0.8 \end{pmatrix}$$

with  $m_2(\underline{e}) = 1.604$ .

**Section 8.7***Exercise 1*

$$r(BA) = \text{dimension of } B(A(R^n)),$$

and from the dimension theorem,

$$\text{dimension of } B(A(R^n)) = \text{dimension of } A(R^n) - \text{dimension of} \\ \text{kernel of } B \text{ with domain } A(R^n).$$

Since the dimension of any vector space is greater than or equal to zero,

$$r(BA) = \text{dimension of } B(A(R^n)) \leq \text{dimension of } A(R^n) = r(A).$$

Similarly,  $r(BA) \leq r(B)$  as well; so the rank of the product of two matrices is less than or equal to the rank of either matrix.



*Exercise 2*

- (i) Equations (2) have solution vector

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 10 \\ -3 \end{pmatrix}$$

Equations (3) have solution vector

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

- (ii) If the coefficients of
- $x_1$
- and
- $x_2$
- in Equations (1) are known to be exact, the solution is valid.

Equations (2) and Equations (3) can be obtained from Equations (1) by making small changes in the  $x_2$  coefficients, amounting to relative changes of  $-0.001$  and  $+0.0001$  respectively. Since the solution vectors for the three systems of equations are markedly different (even the signs change) we conclude that Equations (1) are ill-conditioned.

*Exercise 3*

- (i) The matrix of the iteration process is

$$\begin{pmatrix} 0 & -0.2 & -0.2 \\ -0.1 & 0 & -0.1 \\ -0.2 & -0.2 & 0 \end{pmatrix}$$

The largest column sum of the moduli of the elements is 0.4.

Since  $0.4 < 1$

the iteration process will converge.

- (ii) Using the guess given

$$x_1^{(1)} = 0.2 - 0.02 - 0.04 = 0.14$$

$$x_2^{(1)} = 0.1 - 0.02 - 0.02 = 0.06$$

$$x_3^{(1)} = 0.2 - 0.04 - 0.02 = 0.14$$

$$x_1^{(2)} = 0.2 - 0.012 - 0.028 = 0.16$$

$$x_2^{(2)} = 0.1 - 0.014 - 0.014 = 0.072$$

$$x_3^{(2)} = 0.2 - 0.028 - 0.012 = 0.16$$



$$x_1^{(3)} = 0.2 - 0.0144 - 0.032 = 0.1536$$

$$x_2^{(3)} = 0.1 - 0.016 - 0.016 = 0.068$$

$$x_3^{(3)} = 0.2 - 0.032 - 0.0144 = 0.1536$$

$$x_1^{(4)} = 0.2 - 0.0136 - 0.03072 = 0.1557$$

$$x_2^{(4)} = 0.1 - 0.015 - 0.015 = 0.070$$

$$x_3^{(4)} = 0.2 - 0.03072 - 0.0136 = 0.1557$$

$$x_1^{(5)} = 0.2 - 0.014 - 0.0311 = 0.1549 = 0.155$$

$$x_2^{(5)} = 0.1 - 0.0156 - 0.0156 = 0.0688 = 0.069$$

$$x_3^{(5)} = 0.2 - 0.0311 - 0.014 = 0.1549 = 0.155$$

$$x_1^{(6)} = 0.2 - 0.01376 - 0.03098 = 0.15526 = 0.155$$

$$x_2^{(6)} = 0.1 - 0.0155 - 0.01555 = 0.0689 = 0.069$$

$$x_3^{(6)} = 0.2 - 0.01376 - 0.03098 = 0.15526 = 0.155$$

Hence to three decimal places

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0.155 \\ 0.069 \\ 0.155 \end{pmatrix}$$

Notice that  $x_1$  calculations are identical to those for  $x_3$  and hence we can save calculation time.

#### Exercise 4

$$(R_1) \quad 2x_1 - 3x_2 - 4x_3 = -1$$

$$(R_2) \quad x_1 + 3x_2 + 2x_3 = 2$$

$$(R_3) \quad -3x_1 + 6x_2 - 2x_3 = 0$$

The following is just one of many possible sequences of operations.

$$R_2 \longmapsto R_2 - \frac{1}{2}R_1$$

$$(R_2) \quad \frac{9}{2}x_2 + 4x_3 = \frac{5}{2}$$

$$R_3 \longmapsto R_3 + \frac{3}{2}R_1$$



$$(R_3) \quad \frac{3}{2}x_2 - 8x_3 = -\frac{3}{2}$$

$$R_3 \longmapsto R_3 - \frac{1}{3}R_2$$

$$-\frac{28}{3}x_3 = -\frac{7}{3}$$

i.e.

$$x_3 = \frac{1}{4}.$$

By back-substitution

$$x_2 = \frac{2}{3}(\frac{1}{2}) = \frac{1}{3}$$

and

$$x_1 = \frac{1}{2}(-1 + 1 + 1) = \frac{1}{2}.$$

The solution set is  $\{(\frac{1}{2}, \frac{1}{3}, \frac{1}{4})\}$ .

### Exercise 5

$$\begin{array}{ccc|ccc|c} 2 & 1 & 1 & 1 & 0 & 0 & 5 \\ 3 & 2 & 1 & 0 & 1 & 0 & 7 \\ 4 & 2 & 1 & 0 & 0 & 1 & 8 \end{array}$$

$$R_1 \longmapsto \frac{1}{2}R_1$$

$$\begin{array}{ccc|ccc|c} 1 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 0 & 0 & \frac{5}{2} \\ 3 & 2 & 1 & 0 & 1 & 0 & 7 \\ 4 & 2 & 1 & 0 & 0 & 1 & 8 \end{array}$$

$$R_2 \longmapsto R_2 - 3R_1, R_3 \longmapsto R_3 - 4R_1$$

$$\begin{array}{ccc|ccc|c} 1 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 0 & 0 & \frac{5}{2} \\ 0 & \frac{1}{2} & -\frac{1}{2} & -\frac{3}{2} & 1 & 0 & -\frac{1}{2} \\ 0 & 0 & -1 & -2 & 0 & 1 & -2 \end{array}$$

$$R_1 \longmapsto R_1 - R_2 + R_3$$

$$\begin{array}{ccc|ccc|c} 1 & 0 & 0 & 0 & -1 & 1 & 1 \\ 0 & \frac{1}{2} & -\frac{1}{2} & -\frac{3}{2} & 1 & 0 & -\frac{1}{2} \\ 0 & 0 & -1 & -2 & 0 & 1 & -2 \end{array}$$

$$R_2 \longmapsto 2R_2 - R_3, R_3 \longmapsto -R_3$$



$$\begin{array}{ccc|ccc|c} 1 & 0 & 0 & 0 & -1 & 1 & 1 \\ 0 & 1 & 0 & -1 & 2 & -1 & 1 \\ 0 & 0 & 1 & 2 & 0 & -1 & 2 \end{array}$$

Hence the required inverse matrix is

$$\begin{pmatrix} 0 & -1 & 1 \\ -1 & 2 & -1 \\ 2 & 0 & -1 \end{pmatrix}$$



## CHAPTER 9 COMPLEX NUMBERS

### 9.0 Introduction

In order to understand fully what complex numbers are, and why they are important, we need to know a little of their history. The complex number system is a natural generalization of the real number system, and this generalization was anticipated by the early Greek mathematicians. The basic question which faced the ancients was this:

Is there a number which when multiplied by itself gives  $-1$ ?

It was not difficult for them to decide that there was no such number, for they argued, quite rightly, that the square of a positive or negative quantity must always be positive. On the other hand, it was disconcerting for them to have equations which had solutions only if one allowed the existence of  $\sqrt{-1}$ .

Diophantus (c. A.D. 275) was one of the first mathematicians to recognize that the set of real numbers is, in a sense, incomplete. He attempted to solve the apparently reasonable problem of finding the sides of a right-angled triangle of perimeter 12 and area 7. This leads directly to the equation (in modern notation)

$$6x^2 - 43x + 84 = 0,$$

in which  $x$  is the length of one side of the triangle.

This equation has roots which involve the square root of a negative quantity.

The ancient mathematicians interpreted an equation of this kind as representing an impossible occurrence. Pacioli (1494) stated that the equation  $x^2 + c = bx$  *cannot* be solved unless  $b^2 \geq 4c$ , and Cardan (1545) described the equation  $x^4 + 12 = 6x^2$  as being "impossible", referring to the roots of such equations as "fictitious". However, Cardan did use the square root of a negative number in computation in order to divide 10 into two parts whose product is 40, and he found the two parts to be  $5 + \sqrt{-15}$  and  $5 - \sqrt{-15}$ . Gauss first called expressions of this kind "complex numbers".

In these early days complex numbers had a certain mystical quality.



Mathematicians were sure that they did not exist; and yet, if one supposed that they did, then it was possible to solve certain problems very quickly. They were used as a calculating device, but regarded with deep suspicion, and that suspicion is still reflected in the words we use today. We still talk of a “real” number and an “imaginary” number, as if one were more “real” than the other.

The major misunderstanding of all the early mathematicians was that they had not appreciated that mathematics, unlike physics, is not something which exists, waiting for men to discover its intricacies, but it is man's own creation. Thus, “the square root of minus one” exists *if we say that it exists*; it is up to us to attach meaning to the phrase, which should, of course, be consistent with any previous definitions which we wish to include in the system under consideration.

Wallis (1673) seems to have appreciated the point. He stated that the square root of a negative number was thought to imply the impossible, but that the same might also be said of a negative number, although we can easily explain the latter in a physical application:

“These *Imaginary* Quantities (as they are commonly called), arising from the *Supposed* Root of a Negative Square (when they happen), are reputed to imply that the Case proposed is Impossible.

And so indeed it is, as to the first and strict notion of what is proposed. For it is not possible that any Number (Negative or Affirmative) Multiplied into itself can produce (for instance)  $-4$ . Since that Like Signs (whether  $+$  or  $-$ ) will produce  $+$ ; and therefore not  $-4$ .

But it is also Impossible that any Quantity (though not a Supposed Square) can be *Negative*. Since that it is not possible that any *Magnitude* can be *Less than Nothing* or any *Number Fewer than None*.

Yet is not that Supposition (of Negative Quantities,) either Unuseful or Absurd; when rightly understood. And though, as to the bare Algebraick Notation, it import a Quantity less than nothing: Yet, when it comes to a Physical Application, it denotes as Real a Quantity as if the Sign were  $+$ ; but to be interpreted in a contrary sense.”\*

In this chapter we shall re-examine the problem of defining  $\sqrt{-1}$  in the light of our knowledge of sets, mappings and functions.

\* D. E. Smith, *History of Mathematics* Vol. II (Dover Publications, 1958).

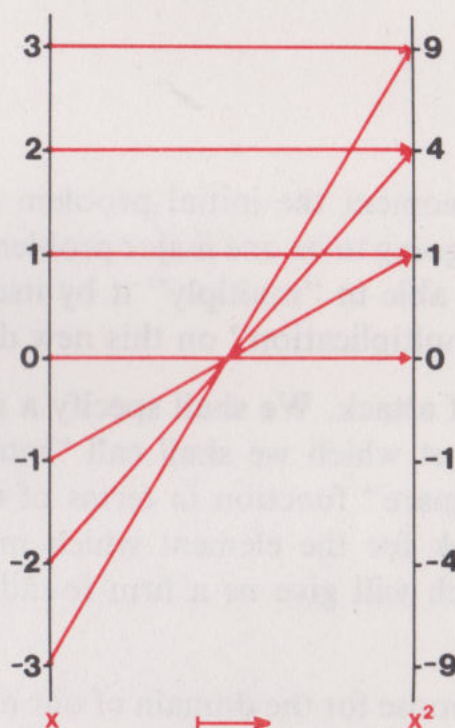


## 9.1 A New “Square” Function

### Introduction

Since we are interested in defining  $\sqrt{\phantom{x}}$ , we shall begin by looking at the “square” function:

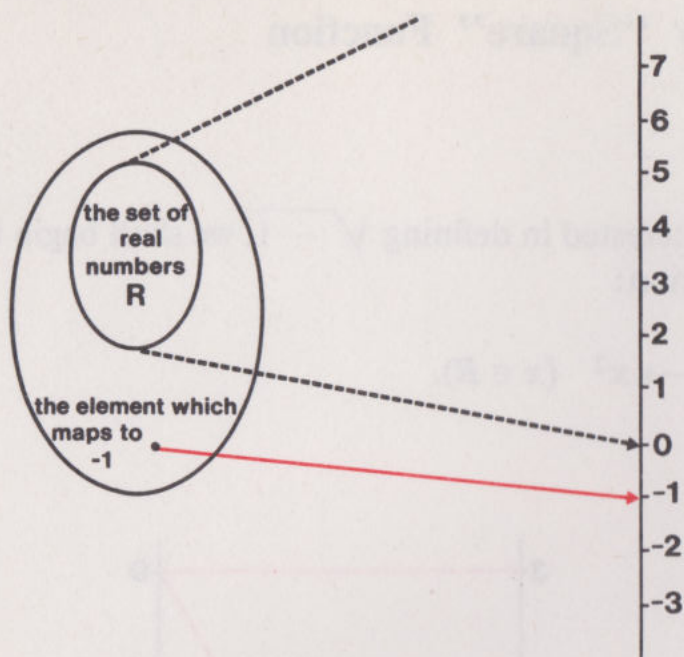
$$f: x \longmapsto x^2 \quad (x \in R).$$



Notice that  $R$  is a suitable codomain for  $f$ , since it contains all the images of  $R$  under  $f$ , although *all* the images are in fact greater than or equal to zero. This is simply another way of stating what the ancients knew: the square of any number is always positive (or zero). The number  $-1$  is certainly not an image of any element in the domain of  $f$ .

We appear to be no further forward, but perhaps it is our definition of the “square” function which is unsatisfactory? Can we enlarge the domain of  $f$  to include an element which maps to  $-1$ ? Can we define a new, more satisfactory “square” function?



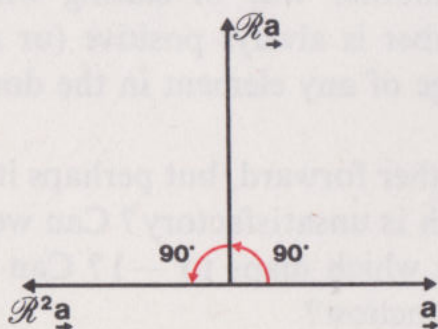


Setting aside for the moment the initial problem of how to specify this larger domain, there is going to be one major problem. In order to “square” something we must be able to “multiply” it by itself. So we are going to need a definition of “multiplication” on this new domain.

This then is our line of attack. We shall specify a new set, and introduce an operation on that set which we shall call “multiplication”. Then we can easily define a “square” function in terms of this operation. Having done that, we can look for the element which maps to  $-1$  under this function. This approach will give us a firm foundation for the study of *complex numbers*.

Which set should we choose for the domain of our new “square” function? The following idea will give us the clue.

Suppose that we start with a geometric vector  $\vec{a}$  and let  $\mathcal{R}$  denote the mapping which “rotates  $\vec{a}$  about its blunt end-point through  $90^\circ$  in an anti-clockwise direction”. This gives us a mapping from the set of geometric vectors to the set of geometric vectors. We write  $\mathcal{R}\vec{a}$  for the result of rotating  $\vec{a}$  through  $90^\circ$  anti-clockwise.





The point to notice is this: applying  $\mathcal{R}$  twice to  $q$  gives  $\mathcal{R} \circ \mathcal{R}q$  which is equal to  $-q$ . In other words, if we write  $\mathcal{R}^2$  for  $\mathcal{R}$  applied twice, then

$$\mathcal{R}^2 q = -q = -1 \times q.$$

We have

$$\mathcal{R}^2 = -I,$$

where  $I$  is the *identity mapping*:

$$I:q \longmapsto q.$$

There are, conceivably, many other sets and mappings on those sets, for which

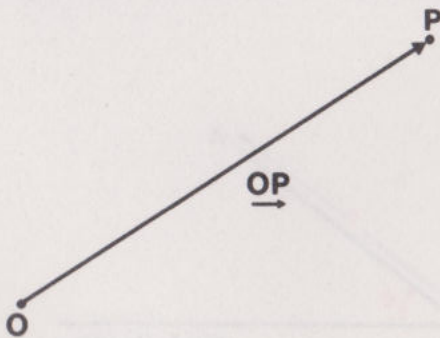
$$(\text{mapping})^2 (\text{element}) = - (\text{element}).$$

We intend to choose the set and the mapping which most suit our needs.

It is important to realize that we have not proved anything; we are just led to an intuitive idea that the set and the mapping which we are looking for may well have something to do with geometric vectors and the idea of rotation.

### The Set of Geometric Vectors

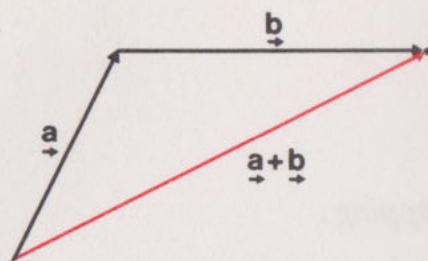
We saw in Chapter 4 that once we specify a fixed point, which we call the origin, then all the points in two or three dimensions can be specified by geometric vectors. If  $O$  denotes the origin, then the geometric vector  $\underline{OP}$  determines the point  $P$ .



We are concerned here with the set of geometric vectors lying in a *plane*; we know that such a set forms a vector space of dimension two, with



suitable definitions of multiplication by a scalar and addition (illustrated in the following diagram). We shall call this set  $V$ .

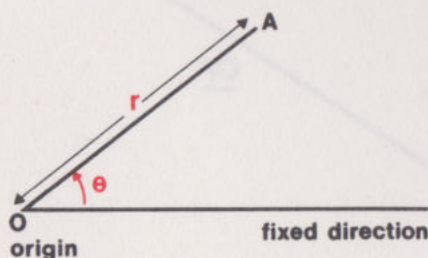


Previously we constructed an interesting example of an “algebra” by introducing a further operation, the inner product. On this occasion we adopt an alternative notion of “multiplication” based on the idea of rotation about  $O$ . It turns out that this new “multiplication” leads us to a very satisfactory algebra with almost all the desirable properties of the algebra of real numbers. Before introducing this new “multiplication” on the set of geometric vectors, we shall need to consider the problem of notation.

### Polar and Cartesian Co-ordinates

Geometric vectors (and also, of course, the points in a plane) can be represented either by *polar co-ordinates*  $(r, \theta)$  or by *Cartesian co-ordinates*  $(x, y)$ : (We shall use red brackets and black brackets to distinguish the two meanings of the number pairs in this chapter. This distinction is not usually made in books but it may be helpful initially.)

In case you haven’t met polar co-ordinates we shall describe them briefly. To obtain polar co-ordinates, we choose a fixed point, called the origin, and a fixed direction.

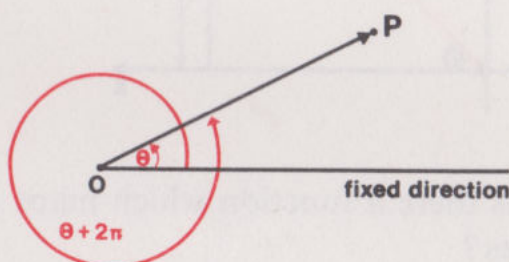


Then given any general point  $A$  in the plane, we can specify its position by the angle  $\theta$  (measured positive in an anti-clockwise direction from the



fixed line) and the distance  $r$  of  $A$  from the origin. Then the numbers  $r$  and  $\theta$  are called **polar co-ordinates** of  $A$ , or of the geometric vector  $\underline{OA}$ . The origin  $O$  has polar co-ordinates  $(0, \theta)$ , where  $\theta$  is arbitrary.

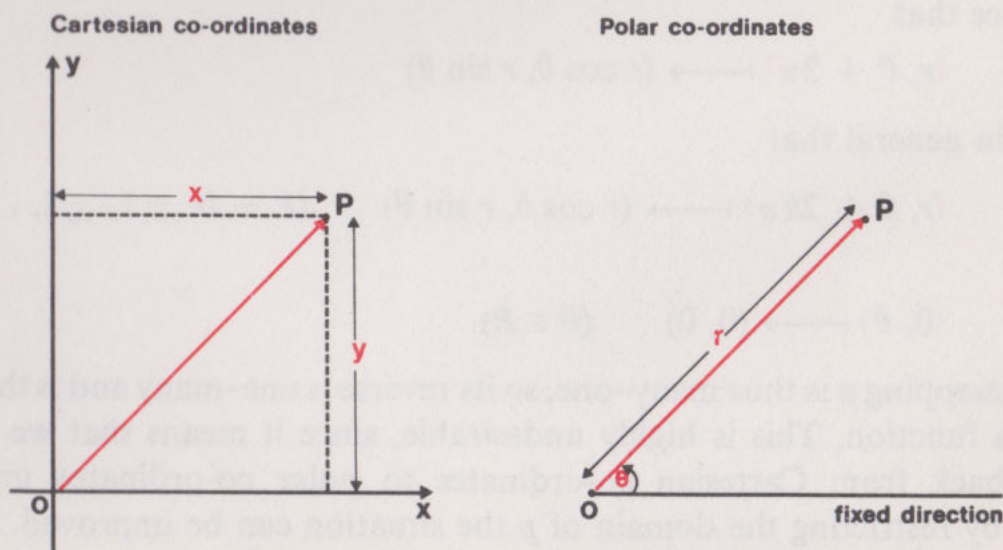
Polar co-ordinates are ideally suited to problems involving rotations and so they seem an obvious choice here. However, polar co-ordinates suffer from two disadvantages. Vector addition is cumbersome in polar co-ordinates, and, if we are given the origin  $O$  and a point  $P$ , then polar co-ordinates of  $P$  are not determined uniquely.



If  $\theta$  gives the direction of  $\underline{OP}$  in radians, then so also do the angles  $\theta + 2k\pi$ ,  $k = \pm 1, \pm 2, \dots$

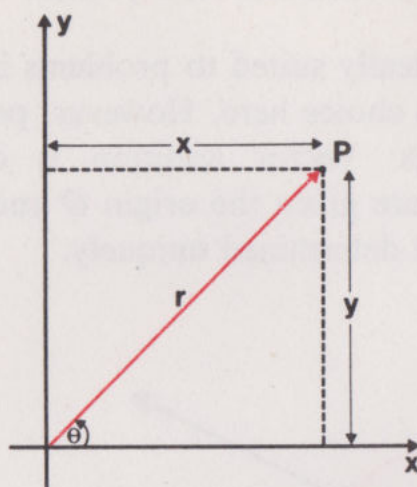
Cartesian co-ordinates do not suffer from these difficulties, but on the other hand they are rather cumbersome when dealing with rotations, as we shall see.

We shall try to get the best from both systems by defining our new “multiplication” in terms of polar co-ordinates; then we shall find the corresponding operation in terms of Cartesian co-ordinates. First we must consider carefully the relationship between the two co-ordinate systems.





If we are given the polar co-ordinates of a point, can we determine its Cartesian co-ordinates, and vice versa?



In terms of mappings: is there a function which maps polar co-ordinates to Cartesian co-ordinates?

If we take the same origin in both cases, and the fixed direction for our polar co-ordinates along the positive  $x$ -axis, then clearly

$$x = r \cos \theta,$$

and

$$y = r \sin \theta.$$

So

$$\begin{array}{ccc} p: (r, \theta) & \longmapsto & (r \cos \theta, r \sin \theta) \\ \text{polar} & & \text{Cartesian} \end{array} \quad ((r, \theta) \in R_0^+ \times R)$$

is the required mapping, and it is indeed a function. ( $R_0^+$  denotes the set of positive real numbers and zero.)

Notice that

$$(r, \theta + 2\pi) \longmapsto (r \cos \theta, r \sin \theta)$$

and in general that

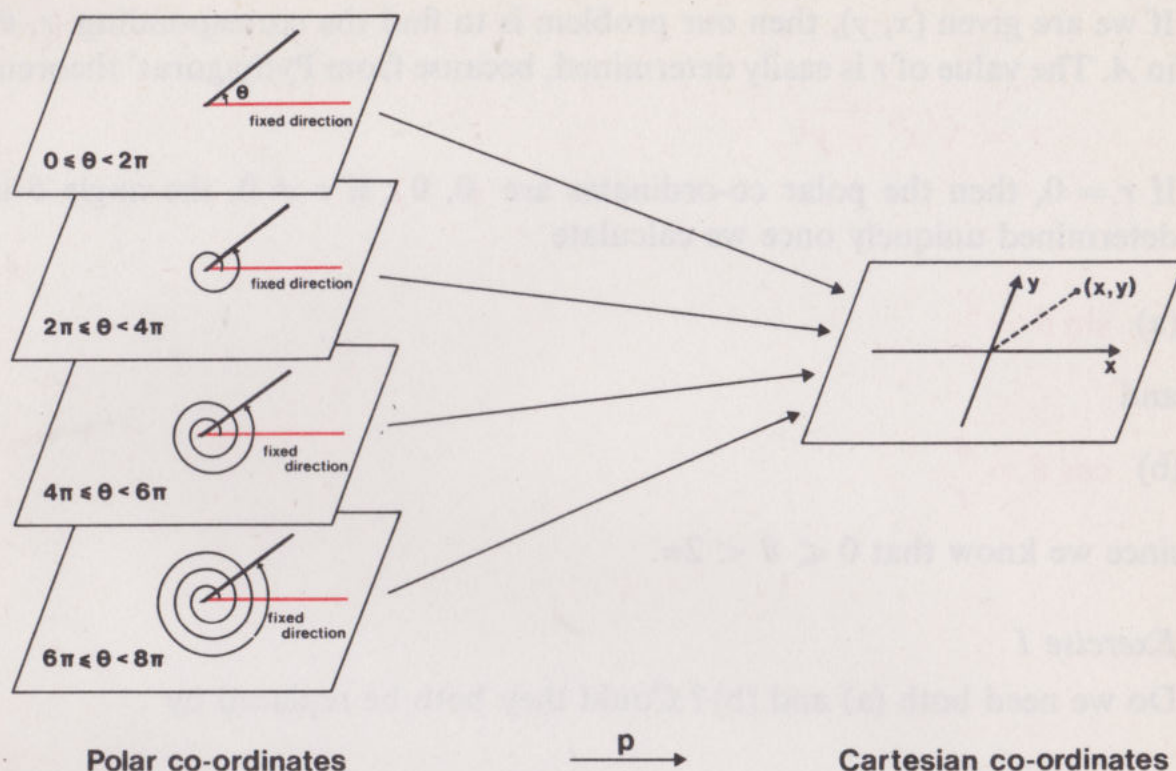
$$(r, \theta + 2k\pi) \longmapsto (r \cos \theta, r \sin \theta) \quad (k = 0, \pm 1, \pm 2, \dots).$$

Also

$$(0, \theta) \longmapsto (0, 0) \quad (\theta \in R).$$

The mapping  $p$  is thus many-one, so its reverse is one-many and is therefore *not* a function. This is highly undesirable, since it means that we cannot get back from Cartesian co-ordinates to polar co-ordinates uniquely. But by restricting the domain of  $p$  the situation can be improved.





The function  $p$  is many-one

Instead of taking  $R_0^+ \times R$  as the domain, we take the set

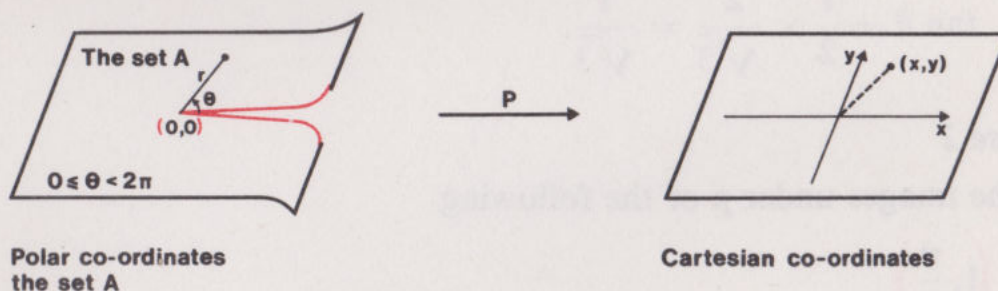
$$A = \{(r, \theta) : r \in R^+, 0 \leq \theta < 2\pi\} \cup \{(0, 0)\}.$$

On this restricted domain,  $p$  is now one-one. We have cut out all the  $(0, \theta)$ 's except  $(0, 0)$  and we have cut out all the  $\theta + 2k\pi$ 's except  $\theta$  itself.

Of course, changing the domain of  $p$  changes  $p$ : the domain is an integral part of the function. We shall use a capital  $P$  for this new function:

$$P: (r, \theta) \longmapsto (r \cos \theta, r \sin \theta) \quad ((r, \theta) \in R_+ \times [0, 2\pi[).^*$$

$$P: (0, 0) \longmapsto (0, 0).$$



The function  $P$  is one-one

The reverse mapping,  $P^{-1}$ , is now also a function.

\*  $[0, 2\pi[$  means the interval  $[0, 2\pi]$  with the end-point at 2 omitted, i.e.

$[0, 2\pi[ = \{\theta : 0 \leq \theta < 2\pi\}.$



If we are given  $(x, y)$ , then our problem is to find the corresponding  $(r, \theta)$  in  $A$ . The value of  $r$  is easily determined, because from Pythagoras' theorem

$$r = \sqrt{x^2 + y^2}.$$

If  $r = 0$ , then the polar co-ordinates are  $(0, 0)$ ; if  $r \neq 0$ , the angle  $\theta$  is determined uniquely once we calculate

$$(a) \quad \sin \theta = \frac{y}{r}$$

and

$$(b) \quad \cos \theta = \frac{x}{r}$$

since we know that  $0 \leq \theta < 2\pi$ .

### Exercise 1

Do we need both (a) and (b)? Could they both be replaced by

$$\tan \theta = \frac{y}{x}$$

when  $x \neq 0$ ?

(i) Find two angles  $\theta$ , such that  $0 \leq \theta < 2\pi$ , for which

$$\sin \theta = \frac{1}{2}.$$

(ii) Find two angles  $\theta$ , such that  $0 \leq \theta < 2\pi$ , for which

$$\cos \theta = \frac{\sqrt{3}}{2}.$$

(iii) Write down the single angle  $\theta$ , such that  $0 \leq \theta < 2\pi$ , for which

$$\sin \theta = \frac{1}{2} \text{ and } \cos \theta = \frac{\sqrt{3}}{2}.$$

(iv) Find the two angles  $\theta$ , such that  $0 \leq \theta < 2\pi$ , for which

$$\tan \theta = \frac{1}{2} \times \frac{2}{\sqrt{3}} = \frac{1}{\sqrt{3}}.$$

### Exercise 2

Find the images under  $p$  of the following

$$(i) \quad \left(1, \frac{\pi}{4}\right)$$

$$(ii) \quad \left(2, \frac{\pi}{3}\right)$$

What are the images of these elements under  $P$ ?



**Exercise 3**

Find the images of the following under the reverse of  $p$ .

- (i)  $(1, \sqrt{3})$
- (ii)  $(\sqrt{2}, -\sqrt{2})$
- (iii)  $(0, 0)$
- (iv)  $(0, 1)$

What are the images of these elements under the reverse of  $P$ ?

**9.2 A New Operation on the Set of Geometric Vectors**

In this section we shall define a new “multiplication” operation  $\circ$  on the set  $A$ , and then we shall have a look at its geometric interpretation.

We arrive at the definition in two stages. First we define such an operation on the set of all polar co-ordinates  $R_0^+ \times R$ . If  $(r_1, \theta_1)$  and  $(r_2, \theta_2)$  are any two elements of this set, then we define  $\circ$  by

$$(r_1, \theta_1) \circ (r_2, \theta_2) = (r_1 r_2, \theta_1 + \theta_2).$$

If  $(r_1, \theta_1)$  and  $(r_2, \theta_2)$  are any two elements of the subset  $A$  of  $R_0^+ \times R$ , then this definition still defines a binary operation on  $A$ , but the operation is not closed. For instance,

$$(1, \pi) \circ \left(3, \frac{3\pi}{2}\right) = \left(3, \frac{5\pi}{2}\right),$$

and the latter is not an element of  $A$ . The non-closure can be a nuisance, because we shall want to restrict our attention *as much as possible* to  $A$  and the function  $P$  (as opposed to  $R_0^+ \times R$  and  $p$ ). So we now define  $\circ$  on  $A$  by

$$(r_1, \theta_1) \circ (r_2, \theta_2) = (r_1 r_2, \theta_1 + \theta_2 \pmod{2\pi}), \quad (r_1, r_2 \neq 0),$$

where  $\theta_1 + \theta_2 \pmod{2\pi}$  means addition modulo  $2\pi$ .

For example, if

$$4\pi > \theta_1 + \theta_2 \geq 2\pi,$$

then

$$\theta_1 + \theta_2 \pmod{2\pi} = \theta_1 + \theta_2 - 2\pi.$$

In general

$$\theta_1 + \theta_2 \pmod{2\pi} = \theta_1 + \theta_2 - 2k\pi$$



where  $k$  is an integer so chosen that

$$0 \leq \theta_1 + \theta_2 \pmod{2\pi} < 2\pi.$$

(We could write  $\theta_1 \oplus_{2\pi} \theta_2$  instead of  $\theta_1 + \theta_2 \pmod{2\pi}$ : see, for instance, Chapter 2 for a definition of  $\oplus_5$ , the operation of addition modulo 5.)

We have not yet overcome the problem of closure, since, by definition,

$$A = \{(r, \theta) : r \in R^+, 0 \leq \theta < 2\pi\} \cup \{(0, 0)\},$$

and our definition of  $\circ$  only applies to the set  $A_1$ , where

$$A_1 = \{(r, \theta) : r \in R^+, 0 \leq \theta < 2\pi\}.$$

So we define

$$(r, \theta) \circ (0, 0) = (0, 0) \circ (r, \theta) = (0, 0) \quad ((r, \theta) \in R^+ \times R)$$

and

$$(0, 0) \circ (0, 0) = (0, 0).$$

The operation  $\circ$  is now a closed binary operation on  $A$ .

We use the same symbol for the binary operations on  $A$  and on  $R_0^+ \times R$ , because geometrically, say in terms of the combination of geometric vectors specified by the polar co-ordinates, the operations are the same. In fact we shall speak of one binary operation, leaving the context to make it clear, if necessary, in which set we are working. Notice that both binary operations are commutative.

Now let us have a look at the geometric interpretation of  $\circ$ .

Geometrically, this operation can be interpreted as follows: take the geometric vector determined by  $(r_1, \theta_1)$ , scale it up (or down) by a factor  $r_2$ , and rotate it about its blunt end-point through an angle  $\theta_2$  anti-clockwise.

Alternatively, since the operation is commutative, we can say: take the geometric vector determined by  $(r_2, \theta_2)$ , scale it up (or down) by a factor  $r_1$ , and rotate it about its blunt end-point through an angle  $\theta_1$  anti-clockwise.





Another interpretation is the following. If we regard  $(r_1, \theta_1)$  as determining a scaling by a factor  $r_1$  and a rotation through an angle  $\theta_1$  (as described above), then we can regard it as determining a function which maps the set of all geometric vectors to itself. (Compare Chapter 4, where we regarded a geometric vector as determining a translation.) In particular, the function determined by  $(r_1, \theta_1)$  maps the geometric vector determined by  $(1, 0)$  to the geometric vector determined by  $(r_1, \theta_1)$ . Then we can regard  $(r_1, \theta_1) \circ (r_2, \theta_2)$  as the composition of the corresponding two functions.

The interesting thing, from our point of view, is the interpretation of the operation, which we have introduced in  $R_0^+ \times R$  or  $A$ , in terms of the corresponding Cartesian co-ordinates. We have a mapping to get us from polar to Cartesian co-ordinates. We now want to turn this into a morphism by selecting the right operation in the image set.

Let  $(r_1, \theta_1)$  and  $(r_2, \theta_2)$  be elements of  $A$  where  $r_1 \neq 0$ ,  $r_2 \neq 0$ . Then

$$P:(r_1, \theta_1) \longmapsto (r_1 \cos \theta_1, r_1 \sin \theta_1) = (x_1, y_1)$$

$$P:(r_2, \theta_2) \longmapsto (r_2 \cos \theta_2, r_2 \sin \theta_2) = (x_2, y_2).$$

Now

$$(r_1, \theta_1) \circ (r_2, \theta_2) = (r_1 r_2, \theta_1 + \theta_2 \pmod{2\pi}).$$

So we define the combination of  $(x_1, y_1)$  and  $(x_2, y_2)$  to correspond to the combination of  $(r_1, \theta_1)$  and  $(r_2, \theta_2)$ ; i.e. if we denote the operation on the image set by  $\otimes$ , then

$$\begin{aligned} (x_1, y_1) \otimes (x_2, y_2) &= P((r_1 r_2, \theta_1 + \theta_2 \pmod{2\pi})) \\ &= (r_1 r_2 \cos(\theta_1 + \theta_2), r_1 r_2 \sin(\theta_1 + \theta_2)). \end{aligned}$$

(Notice that we can drop the  $\pmod{2\pi}$  once we take sines and cosines.)

That is not a very useful result: we would like the right-hand side to be expressed in terms of  $x_1, x_2, y_1$  and  $y_2$ . Notice first that

$$\begin{aligned} r_1 r_2 \cos(\theta_1 + \theta_2) &= r_1 r_2 (\cos \theta_1 \cos \theta_2 - \sin \theta_1 \sin \theta_2) \\ &= x_1 x_2 - y_1 y_2. \end{aligned}$$

Similarly,

$$\begin{aligned} r_1 r_2 \sin(\theta_1 + \theta_2) &= r_1 r_2 (\sin \theta_1 \cos \theta_2 + \cos \theta_1 \sin \theta_2) \\ &= y_1 x_2 + x_1 y_2. \end{aligned}$$

It follows that the corresponding operation on the set of Cartesian co-ordinates, denoted by  $\otimes$ , is defined by

$$(x_1, y_1) \otimes (x_2, y_2) = (x_1 x_2 - y_1 y_2, y_1 x_2 + x_1 y_2)$$

This definition also covers the case when  $(x_1, y_1)$  or  $(x_2, y_2)$  is  $(0, 0)$ .



This formula is very important, but luckily we do not have to remember it. In section 9.4 we shall introduce a very useful notation which enables us to work out “products” quickly and easily.

Notice that  $\circ$  is a simpler operation to perform than  $\otimes$ : this means that polar co-ordinates are easier to use when “multiplying”.

### Summary

In this section we have defined a multiplication operation  $\circ$  on the set of polar co-ordinates; this operation is based on the ideas of scaling and rotation of geometric vectors. We have also found the induced operation  $\otimes$  on the set of Cartesian co-ordinates which corresponds to  $\circ$  on the set of polar co-ordinates under the morphism  $P$ .

We can summarize the way we obtained  $\otimes$  from  $\circ$  by drawing the commutative diagram for the morphism  $P$  for the case  $r_1 \neq 0, r_2 \neq 0$ .

$$\begin{array}{ccc}
 ((r_1, \theta_1), (r_2, \theta_2)) & \xrightarrow{\circ} & (r_1 r_2, \theta_1 + \theta_2 \pmod{2\pi}) \\
 \downarrow P & & \downarrow P \\
 ((x_1, y_1), (x_2, y_2)) & \xrightarrow{\otimes} & (x_1 x_2 - y_1 y_2, y_1 x_2 + x_1 y_2)
 \end{array}$$

### Exercise 1

Fill in the gaps in the following diagrams.

(i)  $\left( \left( 1, \frac{\pi}{4} \right), \left( \frac{1}{2}, \frac{\pi}{3} \right) \right)$

$\downarrow P$

$$\left( \quad, \quad \right) \xrightarrow{\otimes} \left( \quad, \quad \right)$$

(ii)  $\left( \left( 1, \frac{\pi}{4} \right), \left( \frac{1}{2}, \frac{\pi}{3} \right) \right) \xrightarrow{\circ} \left( \quad, \quad \right)$

$\downarrow P$

$$\left( \quad, \quad \right)$$

The final answers to parts (i) and (ii) should be the same.



$$(iii) ((-\sqrt{3}, 1), (-2, -2))$$



$$\left( \begin{array}{c} \phantom{0} \\ \phantom{0} \end{array} , \begin{array}{c} \phantom{0} \\ \phantom{0} \end{array} \right) \xrightarrow{\circ} \left( \begin{array}{c} \phantom{0} \\ \phantom{0} \end{array} , \begin{array}{c} \phantom{0} \\ \phantom{0} \end{array} \right)$$

$$(iv) ((-\sqrt{3}, 1), (-2, -2)) \xrightarrow{\otimes} \left( \begin{array}{c} \phantom{0} \\ \phantom{0} \end{array} , \begin{array}{c} \phantom{0} \\ \phantom{0} \end{array} \right)$$



$$\left( \begin{array}{c} \phantom{0} \\ \phantom{0} \end{array} , \begin{array}{c} \phantom{0} \\ \phantom{0} \end{array} \right)$$

The final answers to parts (iii) and (iv) should be the same.

### 9.3 The Argument

We know that the mapping

$$p: (r, \theta) \longmapsto (x, y) \quad ((r, \theta) \in R_0^+ \times R)$$

is many-one. For a given value of  $r \neq 0$  there are many different angles  $\theta$  which will map to the same pair  $(x, y)$ . Each of these angles is called a *value of the argument* of  $(x, y)$ . The **argument** is itself the set of *all* such values, so that

$$\textcolor{red}{\arg(x, y)}$$

is the set

$$\{\theta: p((r, \theta)) = (x, y)\}.$$

If  $\theta_1$  is any particular angle lying in this set, then

$$\arg(x, y) = \{\theta_1 + 2k\pi, k \in \mathbb{Z}\}.$$

(The argument is sometimes called the *amplitude*.)

$$\text{If } P: (r, \theta) \longmapsto (x, y), \quad (r \neq 0)$$

then we say that the **principal value** of the argument of  $(x, y)$  is  $\theta$ , and we denote it by **Arg**  $(x, y)$ , so

$$\textcolor{red}{\text{Arg}(x, y)} = \theta.$$

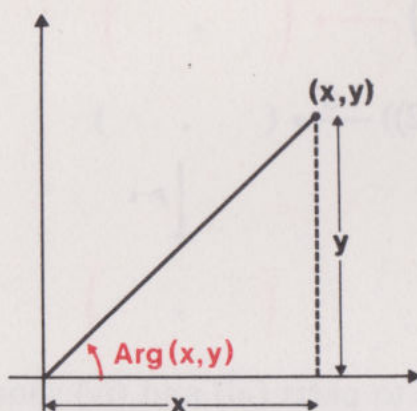
We can regard the principal value of the argument as defining a function

$$\text{Arg}: (x, y) \longmapsto \theta \quad ((x, y) \in R \times R, (x, y) \neq (0, 0)).^*$$

\* Actually we should write  $\text{Arg}(x, y)$  as  $\text{Arg}((x, y))$  but we shall not do this as the meaning is clear.



Notice that  $\text{Arg}(x, y)$  is simply the element of  $\arg(x, y)$  which lies in the interval  $[0, 2\pi[$ , and  $P^{-1}$  picks out this element from the set of all possible angles in the set  $\arg(x, y)$ .



We have seen in Chapter 2 that any many-one mapping defines an equivalence relation on its domain; this is exactly what has happened with  $p$ . Given an  $(x, y)$ , the set of all  $\theta$ 's for which  $p: (r, \theta) \mapsto (x, y)$  forms an equivalence class\*, and we call this class  $\arg(x, y)$ . The mapping  $P^{-1}$  gives us a way of choosing representatives from each of the equivalence classes.

### Exercise 1

Find the argument and its principal value for each of the following:

- (i)  $(1, \sqrt{3})$
- (ii)  $(\sqrt{2}, -\sqrt{2})$
- (iii)  $(0, 1)$

(You will be able to use the results of Exercise 9.1.3.)

The next question is fairly natural. We have defined a function  $\text{Arg}$  on a set on which we have a binary operation  $\otimes$ . Is there an operation  $\square$  such that

$$\text{Arg}((x_1, y_1) \otimes (x_2, y_2)) = \text{Arg}(x_1, y_1) \square \text{Arg}(x_2, y_2)?$$

In fact, although  $\arg$  is not a function, we shall ask a more general question: What is the argument of  $(x_1, y_1) \otimes (x_2, y_2)$  in terms of  $\arg(x_1, y_1)$  and  $\arg(x_2, y_2)$ ? (We assume that these arguments exist.)

\* Actually the equivalence class is the set of pairs  $(r, \theta)$ , but since the  $r$  in each pair in the equivalence class is the same, we have allowed ourselves mathematical licence.



Suppose that we have the following situation:

Polar  
Co-ordinates

Cartesian  
Co-ordinates

$$(r_1, \theta_1) \xrightarrow{p} (x_1, y_1)$$

$$(r_2, \theta_2) \xrightarrow{p} (x_2, y_2)$$

i.e.

$$(r_1 r_2, \theta_1 + \theta_2) \xrightarrow{p} (x_1, y_1) \otimes (x_2, y_2)$$

We therefore know that if  $r_1 \neq 0$  and  $r_2 \neq 0$ , then

$$\arg(x_1, y_1) = \{\theta_1 + 2k\pi, k \in \mathbb{Z}\},$$

$$\arg(x_2, y_2) = \{\theta_2 + 2k\pi, k \in \mathbb{Z}\}$$

and also

$$\arg((x_1, y_1) \otimes (x_2, y_2)) = \{\theta_1 + \theta_2 + 2k\pi, k \in \mathbb{Z}\}.$$

It follows that we can obtain  $\arg((x_1, y_1) \otimes (x_2, y_2))$  from  $\arg(x_1, y_1)$  and  $\arg(x_2, y_2)$  by a sort of addition: we can, for instance, add each of the elements of  $\arg(x_1, y_1)$  to each of the elements of  $\arg(x_2, y_2)$ . It would be simpler to add *one* element of  $\arg(x_1, y_1)$  to each of the elements of  $\arg(x_2, y_2)$  or vice versa.

### Exercise 2

If  $(x_1, y_1) \neq (0, 0)$  and  $(x_2, y_2) \neq (0, 0)$ , what is  $\square$  in

$$\text{Arg}((x_1, y_1) \otimes (x_2, y_2)) = \text{Arg}(x_1, y_1) \square \text{Arg}(x_2, y_2)?$$

### Summary

In this section we have defined the mappings  $\text{Arg}$  and  $\arg$  which respectively associate an angle and a set of angles with the pair of Cartesian co-ordinates  $(x, y)$ .  $\text{Arg}(x, y)$  is the angle (measured anti-clockwise) between the positive  $x$ -axis and the straight line from the origin to the point  $(x, y)$ . We have also found the induced operation  $\square$  on the image set of  $\text{Arg}$  which corresponds to the operation  $\otimes$  on the set of Cartesian co-ordinates without  $(0, 0)$ .

## 9.4 Real and Complex Numbers

Let us now return to the problem we posed in the Introduction to the chapter: finding a new domain for the “square” function, so that we can find an element in the domain which maps to  $-1$ .

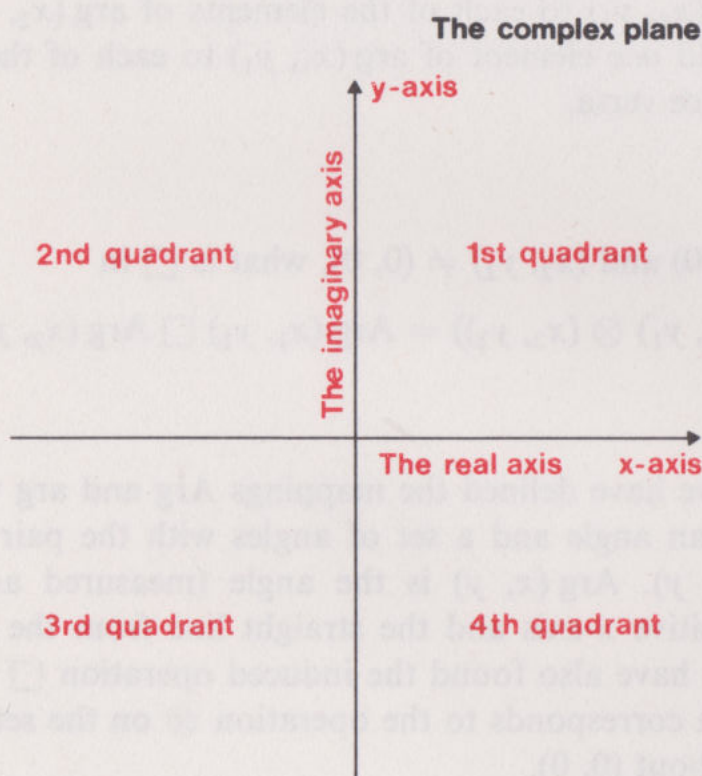


The set which we shall take as the new domain of our “square” function is the set of pairs of real numbers  $(x, y)$ . We denote\* this set by  $C$ ; for the operation of “multiplication” on  $C$  we shall use  $\otimes$ .

We shall call the elements of  $C$  **complex numbers**; that is, each complex number is in fact an ordered pair of real numbers. In order to distinguish the two numbers in this pair (and for historical reasons), we call the first number the **real part** and the second number the **imaginary part** of the complex number.

It is often useful to plot complex numbers  $(x, y)$  on a graph in the usual way. Such a graph in this context is called an **Argand† diagram**.

On an Argand diagram, the set  $\{(x, 0), x \in R\}$  is represented by the  $x$ -axis, and, since we shall identify this set with the set of real numbers, this line is often called the **real axis**. The  $y$ -axis, which represents the points  $\{(0, y), y \in R\}$ , is often called the **imaginary axis**, and the set of points representing  $C$  is often called the **complex plane**.



\*  $C$  is, of course,  $R \times R$ , so why another name? The reason is that ordered pairs of numbers are used in many contexts. We have a particular context here, so we use a symbol for the set of ordered pairs which automatically indicates the context.

† This graphical representation was apparently suggested by the Norwegian surveyor Casper Wessel (1797) and later by several authors including J. R. Argand (1806) and Gauss.



We define addition on  $C$  exactly as we did in Chapter 5 for the vector space formed from  $R \times R$ :

$$(x_1, y_1) + (x_2, y_2) = (x_1 + x_2, y_1 + y_2).$$

Although our definition of addition arose from considering geometric vectors, we now concentrate our attention on the set  $C$  with its algebraic operations of addition and “multiplication”.

One of the requirements of our original discussion was that the new domain of the “square” function should contain  $R$  as a subset. Strictly speaking,  $R$  is not a subset of  $C$ , but consider the subset

$$D = \{(x, 0), x \in R\}.$$

Addition in  $D$  takes the form

$$(x_1, 0) + (x_2, 0) = (x_1 + x_2, 0),$$

and this corresponds exactly to our normal addition of real numbers. If we define a one-one function:

$$f: (x, 0) \longmapsto x \quad ((x, 0) \in D),$$

then we can draw the following commutative diagram.

$$\begin{array}{ccc} ((x_1, 0), (x_2, 0)) & \xrightarrow{+ \text{ (in } C)} & (x_1 + x_2, 0) \\ \textcolor{red}{f} \downarrow & & \downarrow f \\ (x_1, x_2) & \xrightarrow{\textcolor{red}{+} \text{ (in } R)} & x_1 + x_2 \end{array}$$

Since  $(x_1, 0) \otimes (x_2, 0) = (x_1 x_2, 0)$ , we also have

$$\begin{array}{ccc} ((x_1, 0), (x_2, 0)) & \xrightarrow{\otimes} & (x_1 x_2, 0) \\ \textcolor{red}{f} \downarrow & & \downarrow f \\ (x_1, x_2) & \xrightarrow{\textcolor{red}{\times}} & x_1 \times x_2 \end{array}$$

So  $f$  is an isomorphism from  $D$  to  $R$  for both the addition and multiplication operations. Although we do not have  $R$  as a subset of our new domain  $C$ , we do have  $D$  in which the arithmetic is the same as the arithmetic in  $R$ . It is not difficult to rephrase problems in  $R$  in terms of  $D$ . Since  $(-1, 0)$  in  $C$  corresponds to  $-1$  in  $R$ , we can now rephrase our original problem in the following form:

Is there an element  $(x, y) \in C$  such that

$$(x, y) \otimes (x, y) = (-1, 0)?$$

You may now be able to solve this problem; we discuss it below.



### Summary

We have defined the set of complex numbers to be the set  $R \times R$ ; we have also defined the operations of addition (+) and “multiplication” ( $\otimes$ ) on this set. The set  $R$  can be identified with the  $x$ -axis in the complex plane, that is, the set  $D$ . We have an isomorphism:

$$f: (D, +, \otimes) \longmapsto (R, +, \times).$$

### The “Square” Function

We are now in a position to define our new “square” function:

$$\text{sq}: (x, y) \longmapsto (x, y) \otimes (x, y), \quad ((x, y) \in C).$$

We have written the right-hand expression in this way to emphasize the “square”, but from our definition we know that

$$(x, y) \otimes (x, y) = (x^2 - y^2, 2xy),$$

so that

$$\text{sq}: (x, y) \longmapsto (x^2 - y^2, 2xy), \quad ((x, y) \in C).$$

Notice that sq maps  $C$  to  $C$ , and, although sq does not look much like our well known real “square” function ( $x \longmapsto x^2, x \in R$ ), the two functions do have some interesting and, in fact, vital things in common.

We already know that

$$\text{sq}: (x, 0) \longmapsto (x^2, 0) \quad (x \in R)$$

so the restriction of sq onto the subset  $D$  is almost the same as the real “square” function.

Now for the crucial question. Is there an element of  $C$  which maps to  $(-1, 0)$  under sq? In other words, can we choose  $(x, y)$  in such a way that

$$\text{sq}(x, y) = (-1, 0)?$$

(Remember that we are identifying  $(-1, 0)$  with  $-1$ .) We know that

$$\text{sq}(x, y) = (x^2 - y^2, 2xy)$$

and number pairs can only be equal if the corresponding elements are equal. We require that

$$(x^2 - y^2, 2xy) = (-1, 0),$$

that is,

$$x^2 - y^2 = -1$$

and

$$2xy = 0.$$



The second equation implies that either  $x = 0$  or  $y = 0$ . If  $y = 0$ , then the first equation cannot possibly be true for any real  $x$ . On the other hand,  $x = 0$  implies that  $y = \pm 1$ . We have therefore shown that

$$\text{sq}:(0, 1) \longmapsto (-1, 0)$$

and

$$\text{sq}:(0, -1) \longmapsto (-1, 0).$$

The complex numbers  $(0, -1)$  and  $(0, 1)$  are the “square roots” of  $(-1, 0)$ , and we are gratified to find that there are two “square roots”, just as there are two real square roots of any positive real number.

### A Useful Notation

The development of complex numbers so far in this text has been aimed at giving a firm base on which to build, but the notation which we have used is not very practical. We shall now introduce a notation which is a considerable aid to computation.

We can rewrite the complex number

$$(x, y) = (x, 0) + (0, y)$$

in the form  $x + iy$ , or sometimes  $x + yi$ , where the letter  $i$  is used merely as a notational device to indicate that  $y$  is the second number in the original pair. As we have noted before,  $x$  is called the *real part* of the complex number and  $y$  is called the *imaginary part*.<sup>\*</sup> The rule for “multiplication” of complex numbers:

$$(x_1, y_1) \otimes (x_2, y_2) = (x_1x_2 - y_1y_2, x_1y_2 + x_2y_1)$$

then becomes

$$(x_1 + iy_1)(x_2 + iy_2) = (x_1x_2 - y_1y_2) + i(x_1y_2 + x_2y_1),$$

if we follow the convention of the algebra of real numbers and drop the special symbol  $\otimes$  for “multiplication”. (Occasionally, in the following sections, we shall insert the symbol  $\otimes$  when we wish to emphasize its use.)

You can easily check that if we multiply out the left-hand side of the last equation as for ordinary real algebra, and whenever we encounter  $i \times i$  we replace it by  $-1$ , then we get the right-hand side.

We emphasize that there is no suggestion that  $i$  is some sort of distorted real number; it is simply a device for separating the two parts of a complex

<sup>\*</sup> It is a common mistake to say that the imaginary part is  $iy$  rather than  $y$ .



number. However, we can use ordinary algebraic rules when manipulating elements like  $x + iy$ , and this justifies calling them “complex” numbers. It is common practice to represent  $x + iy$  by  $z$  for convenience. We shall denote the real and imaginary parts of the complex number  $z$  by  $\operatorname{Re} z$  and  $\operatorname{Im} z$  respectively; that is,

$$\operatorname{Re} z = x$$

and

$$\operatorname{Im} z = y.$$

We have already seen the mapping

$$P^{-1}: z \longmapsto (r, \theta) \quad z \in \mathbb{C}$$

(but in the slightly different form  $(x, y) \longmapsto (r, \theta)$ ), where  $r = \sqrt{x^2 + y^2}$  and  $\theta = \operatorname{Arg} z$ .

We know that  $x = r \cos \theta$  and  $y = r \sin \theta$ . If we write

$$z = x + iy = r \cos \theta + ir \sin \theta$$

i.e.  $z = r(\cos \theta + i \sin \theta)$ ,

then we say that the right-hand expression is the **polar form** of the complex number  $z$ .

We shall write  $z \times z$  as  $z^2$ , and so on.

### Exercise 1

- (i) Simplify each of the following into the form  $x + iy$ :
  - (a)  $(1 + i)(2 + i)(2 - i)$
  - (b)  $(1 + 2i)(1 - 2i)$
  - (c)  $1 + 3i(4 + 5i) + (2 - i)(1 + i)$
- (ii) If  $z_1 = 1 + 3i$  and  $z_2 = 2 - i$ , evaluate  $z_1^2 z_2$ .
- (iii) Plot the points corresponding to  $1 + 3i$  and  $2 - i$  on an Argand diagram.

### Exercise 2

Find:

- (i)  $\operatorname{Re}((1 + 2i)^2)$  and  $\operatorname{Im}((1 + 2i)^2)$ ;
- (ii)  $\operatorname{Re}(1 + 2i + 3i^2 + i^3)$  and  $\operatorname{Im}(1 + 2i + 3i^2 + i^3)$ ;
- (iii)  $\operatorname{Arg}(1 + i)$  and  $\operatorname{Arg}((1 + i)^2)$ .



*Exercise 3*

Let  $z_1 = x_1 + iy_1,$

$$z_2 = x_2 + iy_2,$$

$$z_3 = x_3 + iy_3.$$

Show that

$$z_1(z_2 + z_3) = z_1z_2 + z_1z_3$$

and that

$$(z_2 + z_3)z_1 = z_2z_1 + z_3z_1.$$

*Exercise 4*

Let  $z_1 = r_1 (\cos \theta_1 + i \sin \theta_1),$

$$z_2 = r_2 (\cos \theta_2 + i \sin \theta_2),$$

$$z_3 = r_3 (\cos \theta_3 + i \sin \theta_3).$$

Show that

$$z_1(z_2z_3) = (z_1z_2)z_3.$$

## 9.5 Summary of Properties of Complex Numbers

We began with the ancient problem of defining  $\sqrt{-1}$ , and, by extending the domain of the “square” function to the set of complex numbers, we were able to find elements which map to  $(-1, 0)$ . In our new notation we could replace  $(-1, 0)$  by  $-1 + i0$  so that

$$(0 + i)^2 = -1 + i0,$$

and

$$(0 - i)^2 = -1 + i0.$$

Normally we could simplify these expressions still further and write

$$i^2 = -1$$

and

$$(-i)^2 = -1.$$

Such abbreviation suggests the commonly used (but suspect) statement that  $i$  and  $-i$  are the square roots of  $-1$ .

We have in fact done more than simply examine a piece of mathematical history. The system which we have developed is a powerful extension of



the algebra of real numbers. There are many strategic advantages to be gained from extending our number system to include the complex numbers, if only because the real number system is, in a sense, incomplete. For example, it is true that a polynomial equation of degree  $n$  has  $n$  complex solutions (some of which may coincide), but, for instance, the polynomial equation  $x^2 + 1 = 0$  has no real solutions.

Eliminating complex numbers from mathematics and its applications today would have almost as drastic an effect as eliminating the negative numbers.

Some of the properties of complex numbers are listed below.

- (i)  $x + iy$  is simply a convenient way of writing the complex number  $(x, y)$ .
- (ii)  $x_1 + iy_1 = x_2 + iy_2$  if and only if  $x_1 = x_2$  and  $y_1 = y_2$ .
- (iii)  $(x_1 + iy_1) + (x_2 + iy_2) = (x_1 + x_2) + i(y_1 + y_2)$   
(the complex numbers are closed for addition).
- (iv)  $z_1 + z_2 = z_2 + z_1$   
(addition is commutative).
- (v)  $z_1 + (z_2 + z_3) = (z_1 + z_2) + z_3$   
(addition is associative).
- (vi)  $(x_1 + iy_1)(x_2 + iy_2) = (x_1x_2 - y_1y_2) + i(x_1y_2 + x_2y_1)$   
(the complex numbers are closed for multiplication).
- (vii)  $z_1z_2 = z_2z_1$   
(multiplication is commutative).
- (viii)  $z_1(z_2z_3) = (z_1z_2)z_3$   
(multiplication is associative).
- (ix)  $z_1(z_2 + z_3) = z_1z_2 + z_1z_3$   
 $(z_2 + z_3)z_1 = z_2z_1 + z_3z_1$   
(multiplication is distributive over addition).
- (x) There are two complex numbers,  $0 + 0i$  and  $1 + 0i$  (which are not equal), with the properties that, for any complex number  $z$ ,

$$z + (0 + 0i) = z \quad \text{and} \quad z(1 + 0i) = z.$$

Notice particularly that the complex numbers are closed for multiplication. You will recall that we constructed a generalization of "multiplication" in Chapter 4 which we called the *inner product*, and we noticed that it was *not* a closed binary operation on the set of geometric vectors, and therefore it was difficult to define an extension of division adequately. On this occasion that difficulty does not arise.



We have seen effectively three ways of representing a complex number:

- (1) in its Cartesian form  $(x, y)$  (which we often prefer to write as  $z = x + iy$ );
- (2) in polar form  $(r, \theta)$ ;
- (3) as a point on an Argand diagram.

Often we switch from one representation to the other, and, although we ought to distinguish between them, we may sometimes refer to “the point  $x + iy$ ”, or we might say, for example, that “a complex number lies on a straight line drawn between two other complex numbers”. In other words, we use the representation which most suits our purpose, without going into a lengthy explanation each time.

### Exercise 1

Show that

- (i)  $(1 + i)(\frac{1}{2} - \frac{1}{2}i) = 1$
- (ii)  $(3 + 4i)(\frac{3}{25} - \frac{4}{25}i) = 1$
- (iii)  $(a + ib)(a - ib) = a^2 + b^2$ .

(Notice that, as mentioned above, we are abbreviating  $x + 0i$  to  $x$  and therefore we have written 1 and  $a^2 + b^2$  on the right-hand sides.)

### Exercise 2

If  $z = r(\cos \theta + i \sin \theta)$ , where  $r \neq 0$ , what is  $\arg z$ ?

## 9.6 The Algebra of Complex Numbers

### Division

First let us understand clearly what division means on the set  $R$ . Suppose that we know how to multiply real numbers; then for any non-zero real number  $q$  we know that

$$q \times \frac{1}{q} = 1.$$

We can define division by

$$p \div q = p \times \frac{1}{q},$$

where  $p \in R$ .



In fact, whenever we have a closed binary operation  $\circ$  on a set, an (identity) element  $e$  in the set such that

$$e \circ x = x \circ e = x$$

and an element  $\tilde{x}$  such that

$$\tilde{x} \circ x = x \circ \tilde{x} = e$$

for every element  $x$  in the set, then we can define an *inverse operation*  $\tilde{\circ}$  of  $\circ$  by putting

$$a \tilde{\circ} b = a \circ \tilde{b}.$$

For real numbers we could take  $\circ$  to be  $\times$ ,  $e$  to be 1 and  $\tilde{x}$  to be  $\frac{1}{x}$  ( $x \neq 0$ ); then the operation  $\tilde{\circ}$  so defined is simply  $\div$ .

Let us follow exactly the same reasoning: given a complex number  $\alpha = a + ib$ , our first task is to find another complex number  $z = x + iy$  such that  $\alpha z = 1$ .

Much of the following work is given in the form of exercises, since most of it is a straightforward development of concepts you have already met.

### Exercise 1

If

$$(a + ib)(x + iy) = 1,$$

show that

$$x = \frac{a}{a^2 + b^2} \quad \text{and} \quad y = \frac{-b}{a^2 + b^2}.$$

Are there any complex numbers  $a + ib$  for which such a number  $x + iy$  does not exist?

From the last exercise we see that a sensible definition of  $\frac{1}{\alpha}$  is

$$\frac{a - ib}{a^2 + b^2}.$$

provided that  $a + ib \neq 0$ .

We can now define **division** of complex numbers by

$$z_1 \div z_2 = z_1 \otimes \frac{1}{z_2} \quad (z_2 \neq 0).$$

(Instead of  $z_1 \div z_2$ , we often write  $\frac{z_1}{z_2}$ .) This definition leads us to define two further terms in connection with complex numbers.



## Complex Conjugate and Modulus

If

$$z = x + iy \quad (z \neq 0),$$

then we have defined  $\frac{1}{z}$  to be

$$\frac{x - iy}{x^2 + y^2}.$$

If  $z = x + iy$ , then we define  $x - iy$  to be the **conjugate** of  $z$ , which we denote by  $\bar{z}$ . (We read  $\bar{z}$  as “ $z$  bar”.)

i.e.  $\bar{z} = x - iy.$

If  $z = x + iy$ , then we define the positive (or zero) number  $r = \sqrt{x^2 + y^2}$  to be the **modulus** of  $z$  and we denote it by  $|z|$ . (We read  $|z|$  as “mod  $z$ ”.)

i.e.  $|z| = \sqrt{x^2 + y^2}.$

Notice that this is consistent with our definition of the modulus of a real number, for if  $y = 0$ , then  $|z| = \sqrt{x^2} = |x|.$

### Example 1

If  $z_1 = 2 + 3i$ ,  $z_2 = 2 - 3i$ ,  $z_3 = 3$ ,  $z_4 = -3i$ ,

then

$$z_2 = \bar{z}_1$$

$$z_1 = \bar{z}_2$$

$$\bar{z}_3 = 3$$

$$\bar{z}_4 = 3i$$

$$|z_1| = \sqrt{2^2 + 3^2} = \sqrt{13} = |z_2|$$

$$|z_3| = \sqrt{3^2} = 3 = |z_4|.$$

### Exercise 2

Verify the following results given in terms of the new definitions.

(i)  $z\bar{z} = |z|^2$

(ii)  $\frac{1}{z} = \frac{\bar{z}}{|z|^2} \quad (z \neq 0)$

(iii)  $z_1 \div z_2 = z_1 \times \frac{\bar{z}_2}{|z_2|^2} \quad (z_2 \neq 0)$

(iv)  $z + \bar{z} = 2 \operatorname{Re} z$

(v)  $z - \bar{z} = 2i \operatorname{Im} z.$



*Exercise 3*

Reduce each of the following expressions to the form  $x + iy$ :

$$(i) \frac{1-i}{3+i} \quad (ii) \frac{1}{1+i} + \frac{1+i}{i}.$$

We have introduced three functions:

$$\text{Arg}: z \longmapsto \text{Arg } z \quad (z \in C, z \neq 0),$$

$$\text{mod}: z \longmapsto |z| \quad (z \in C),$$

$$\text{conj}: z \longmapsto \bar{z} \quad (z \in C),$$

which turn up frequently in calculations with complex numbers. Their usefulness depends very much on a familiarity with their behaviour with respect to addition and multiplication. So in the following exercises we ask you to look at this.

*Exercise 4*

Which of the following statements are true?

(i) The function *mod* is a morphism of  $(C, +)$  to  $(R_0^+, +)$ ; that is

$$|z_1 + z_2| = |z_1| + |z_2|$$

(ii) The function *mod* is a morphism of  $(C, \otimes)$  to  $(R_0^+, \times)$ ; that is

$$|z_1 \otimes z_2| = |z_1| \times |z_2|$$

*Exercise 5*

Which of the following statements are true?

(i) The function *conj* is a morphism of  $(C, +)$  to  $(C, +)$

(ii) The function *conj* is a morphism of  $(C, \otimes)$  to  $(C, \otimes)$ .

**Summary**

In this section we have defined division by

$$z_1 \div z_2 = \frac{z_1}{z_2} = z_1 \otimes \frac{1}{z_2} = \frac{z_1 \bar{z}_2}{|z_2|^2} \quad (z_2 \neq 0).$$

We have also proved the following results:

$$(i) |z_1 z_2| = |z_1| |z_2| \quad (\text{Exercise 4})$$

$$(ii) \overline{z_1 + z_2} = \bar{z}_1 + \bar{z}_2 \quad (\text{Exercise 5})$$

$$(iii) \overline{z_1 z_2} = \bar{z}_1 \bar{z}_2 \quad (\text{Exercise 5})$$



$$(iv) \quad z\bar{z} = |z|^2 \quad (\text{Exercise 2})$$

$$(v) \quad z + \bar{z} = 2 \operatorname{Re} z \quad (\text{Exercise 2})$$

$$(vi) \quad z - \bar{z} = 2i \operatorname{Im} z \quad (\text{Exercise 2})$$

The following results, although simple, are often useful:

$$(vii) \quad \overline{(\bar{z})} = z$$

$$(viii) \quad (z = 0) \text{ if and only if } (|z| = 0)$$

$$(ix) \quad |z| = |\bar{z}|$$

$$(x) \quad |z| \geq \operatorname{Re} z.$$

### Products in Polar Form

It is often a useful calculating device to convert a complex number  $z = x + iy$  into its polar form\* before multiplying, so that

$$x + iy = r(\cos \theta + i \sin \theta),$$

where  $\theta$  is one value of  $\arg z$ .

This is particularly so when calculating high integer powers of  $z$ .

If  $z_1 = r_1(\cos \theta_1 + i \sin \theta_1)$  and  $z_2 = r_2(\cos \theta_2 + i \sin \theta_2)$ , then we know from our definition of multiplication that

$$z_1 z_2 = r_1 r_2 (\cos(\theta_1 + \theta_2) + i \sin(\theta_1 + \theta_2)).$$

We can now see that if  $z = r(\cos \theta + i \sin \theta)$ , then

$$z^2 = r^2 (\cos 2\theta + i \sin 2\theta)$$

and, in general, for any positive integer  $n$ ,

$$z^n = r^n (\cos n\theta + i \sin n\theta).$$

Notice also that  $z^n = r^n (\cos \theta + i \sin \theta)^n$  so that

$$(\cos \theta + i \sin \theta)^n = (\cos n\theta + i \sin n\theta)$$

for any positive integer  $n$ . This result is a special case of a theorem known as **De Moivre's Theorem**.

#### Example 2

To evaluate  $(1 + i)^{10}$  we first write  $1 + i$  in its polar form

$$\sqrt{2} \left( \cos \frac{\pi}{4} + i \sin \frac{\pi}{4} \right).$$

\* Some authors abbreviate  $\cos \theta + i \sin \theta$  to  $\operatorname{cis} \theta$ .



Then

$$\begin{aligned}(1 + i)^{10} &= (\sqrt{2})^{10} \left( \cos \frac{10\pi}{4} + i \sin \frac{10\pi}{4} \right) \\ &= 32(0 + i) = 32i\end{aligned}$$

### Exercise 6

If  $z = \frac{1}{4}(1 + i\sqrt{3})$ ,

calculate  $|z^n|$  for any positive integer  $n$ .

## 9.7 Additional Exercises

### Exercise 1

Use the fact that

$$(\cos \theta + i \sin \theta)^n = (\cos n\theta + i \sin n\theta)$$

to prove that

$$\cos 3\theta = \cos^3 \theta - 3 \cos \theta \sin^2 \theta.$$

### Exercise 2

- (i) Represent  $z$  and  $\bar{z}$  on an Argand diagram for a general complex number  $z = x + iy$ .
- (ii) Give a geometric interpretation of  $|z_1 - z_2|$ , where  $z_1$  and  $z_2$  are any complex numbers.

### Exercise 3

Prove that

- (i)  $\operatorname{Re}(z) \leq |z|$
- (ii)  $|z_1 + z_2| \leq |z_1| + |z_2|$  (the triangle inequality)  
(HINT: write  $|z_1 + z_2|^2 = (z_1 + z_2)(\overline{z_1 + z_2})$  and use the properties listed in the Summary on page 356.)
- (iii)  $|z_1 - z_2| \geq ||z_1| - |z_2||$ .

### Exercise 4

Show that the solution set of the equation

$$x^2 + 2x + 4 = 0 \quad (x \in \mathbb{R})$$



is empty. Show that, if we rewrite this equation as an equation in  $C$ , i.e.

$$z^2 + (2, 0)z + (4, 0) = (0, 0) \quad (z \in C),$$

which we shall write as

$$z^2 + 2z + 4 = 0 \quad (z \in C),$$

then the solution set is  $\{-1 + i\sqrt{3}, -1 - i\sqrt{3}\}$ .

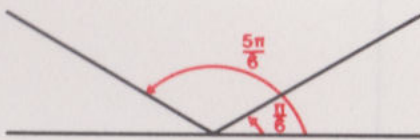
## 9.8 Answers to Exercises

### Section 9.1

#### Exercise 1

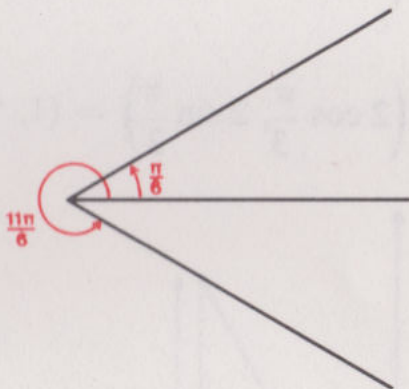
We do indeed require both (a) and (b) and the rest of the exercise is intended to show why.

(i)



Answers:  $\frac{\pi}{6}$  and  $\frac{5\pi}{6}$

(ii)



Answers:  $\frac{\pi}{6}$  and  $\frac{11\pi}{6}$

(iii)  $\frac{\pi}{6}$

(iv)  $\frac{\pi}{6}$  and  $\frac{7\pi}{6}$



In general, any *one* of the equations

$$\sin \theta = a$$

$$\cos \theta = b$$

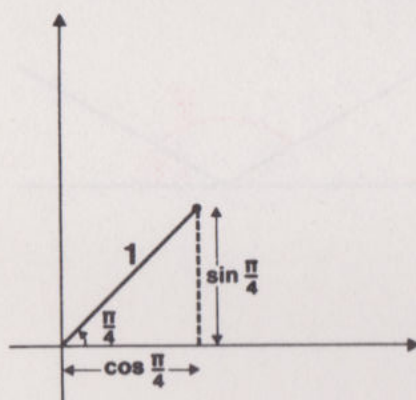
$$\tan \theta = c$$

where  $-1 < a < 1$ ,  $-1 < b < 1$ , will have *two* solutions in the interval  $[0, 2\pi[$  corresponding to *two* points in the plane. For instance, the points with Cartesian co-ordinates  $(x, y)$  and  $(-x, -y)$  both have a polar co-ordinate  $\theta$  satisfying

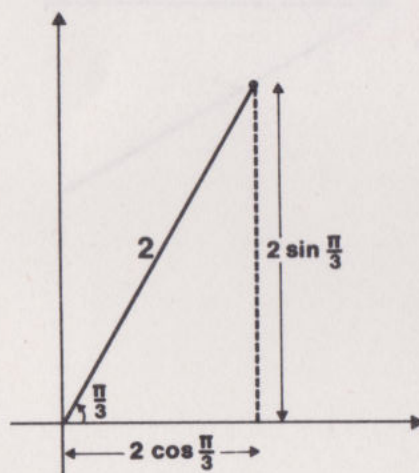
$$\tan \theta = \frac{y}{x} \quad (x \neq 0).$$

### Exercise 2

(i) The image under  $p$  is  $\left(\cos \frac{\pi}{4}, \sin \frac{\pi}{4}\right) = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$ .



(ii) The image under  $p$  is  $\left(2 \cos \frac{\pi}{3}, 2 \sin \frac{\pi}{3}\right) = (1, \sqrt{3})$ .

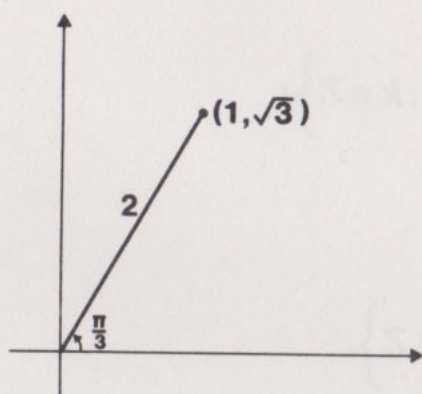


The images under  $P$  are the same for both (i) and (ii).

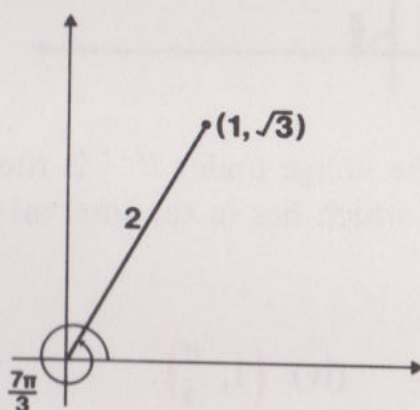


## Exercise 3

(i)

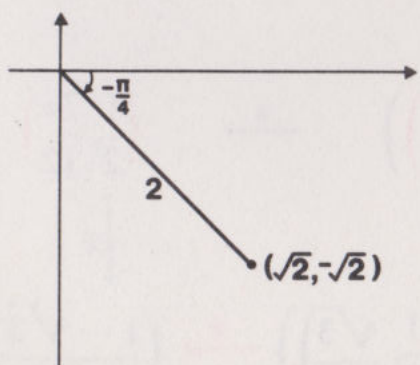


The pair  $\left(2, \frac{\pi}{3}\right)$  is one image of  $(1, \sqrt{3})$  under the reverse of  $p$ , but there are many more, for example,  $\left(2, \frac{7\pi}{3}\right)$ .



In fact there is an infinite set of images. It is  $\left\{\left(2, \frac{\pi}{3} + 2k\pi\right), k \in \mathbb{Z}\right\}$ .

(ii)





You may have answered

$$\left\{ \left( 2, -\frac{\pi}{4} + 2k\pi \right), k \in \mathbb{Z} \right\},$$

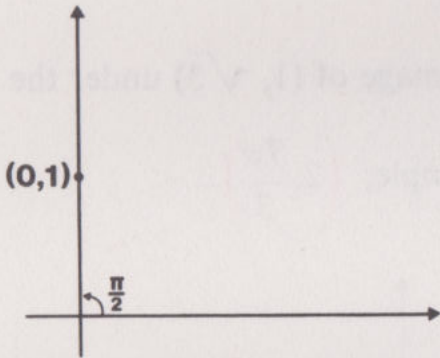
or

$$\left\{ \left( 2, \frac{7\pi}{4} + 2k\pi \right), k \in \mathbb{Z} \right\};$$

both are correct.

(iii)  $\{(0, \theta), \theta \in \mathbb{R}\}.$

(iv)  $\left\{ \left( 1, \frac{\pi}{2} + 2k\pi \right), k \in \mathbb{Z} \right\}.$



In each case except (iii) the image under  $P^{-1}$  is the element of the image set under the reverse of  $p$  which lies in the interval  $[0, 2\pi[$ .

The answers are

(i)  $\left( 2, \frac{\pi}{3} \right)$     (ii)  $\left( 2, \frac{7\pi}{4} \right)$     (iv)  $\left( 1, \frac{\pi}{2} \right).$

In the case of (iii) the image is  $(0, 0).$

Section 9.2

Exercise 1

(i) and (ii)

$$\begin{array}{ccc} \left( \left( 1, \frac{\pi}{4} \right), \left( \frac{1}{2}, \frac{\pi}{3} \right) \right) & \xrightarrow{\circ} & \left( \frac{1}{2}, \frac{7\pi}{12} \right) \\ \downarrow P & & \downarrow P \\ \left( \left( \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right), \left( \frac{1}{4}, \frac{\sqrt{3}}{4} \right) \right) & \xrightarrow{\otimes} & \left( \frac{1 - \sqrt{3}}{4\sqrt{2}}, \frac{1 + \sqrt{3}}{4\sqrt{2}} \right) \end{array}$$



(iii) and (iv)

$$((- \sqrt{3}, 1), (-2, -2)) \xrightarrow{\otimes} (2\sqrt{3} + 2, 2\sqrt{3} - 2)$$

$$\downarrow p^{-1}$$

$$\downarrow p^{-1}$$

$$\left( \left( 2, \frac{5}{6}\pi \right), \left( 2\sqrt{2}, \frac{5}{4}\pi \right) \right) \xrightarrow{\circ} \left( 4\sqrt{2}, \frac{\pi}{12} \right)$$

The more difficult calculation is the one on the extreme right. In (ii) we can write

$$\cos \frac{7}{12}\pi = \cos \left( \frac{\pi}{4} + \frac{\pi}{3} \right) = \cos \frac{\pi}{4} \cos \frac{\pi}{3} - \sin \frac{\pi}{4} \sin \frac{\pi}{3} = \frac{1 - \sqrt{3}}{2\sqrt{2}},$$

$$\sin \frac{7}{12}\pi = \sin \left( \frac{\pi}{4} + \frac{\pi}{3} \right) = \sin \frac{\pi}{4} \cos \frac{\pi}{3} + \cos \frac{\pi}{4} \sin \frac{\pi}{3} = \frac{1 + \sqrt{3}}{2\sqrt{2}}.$$

In (iv) we have to calculate

$$\begin{aligned} r &= \sqrt{(2\sqrt{3} + 2)^2 + (2\sqrt{3} - 2)^2} \\ &= \sqrt{12 + 4 + 8\sqrt{3} + 12 + 4 - 8\sqrt{3}} \\ &= \sqrt{32} = 4\sqrt{2}, \end{aligned}$$

and  $\theta$  from

$$\cos \theta = \frac{2\sqrt{3} + 2}{4\sqrt{2}} = \frac{\sqrt{3} + 1}{2\sqrt{2}}$$

$$\sin \theta = \frac{2\sqrt{3} - 2}{4\sqrt{2}} = \frac{\sqrt{3} - 1}{2\sqrt{2}}$$

These expressions need simplification, conversion to decimal form and then tables to find the angle  $\theta = 15^\circ = \frac{\pi}{12}$ ; alternatively, simplification using trigonometric identities can be used.

### Section 9.3

#### Exercise 1

$$(i) \arg(1, \sqrt{3}) = \left\{ \frac{\pi}{3} + 2k\pi, k \in \mathbb{Z} \right\}$$

$$\text{Arg}(1, \sqrt{3}) = \frac{\pi}{3}$$



$$(ii) \arg(\sqrt{2}, -\sqrt{2}) = \left\{ -\frac{\pi}{4} + 2k\pi, k \in \mathbb{Z} \right\}$$

$$\text{Arg}(\sqrt{2}, -\sqrt{2}) = \frac{7}{4}\pi$$

$$(iii) \arg(0, 1) = \left\{ \frac{\pi}{2} + 2k\pi, k \in \mathbb{Z} \right\}$$

$$\text{Arg}(0, 1) = \frac{\pi}{2}$$

### Exercise 2

Looking back at the table of corresponding polar and Cartesian coordinates, and replacing  $p$  by  $P$ , it is tempting to say that  $\square$  is  $+$ . But this is not quite true. For instance, consider

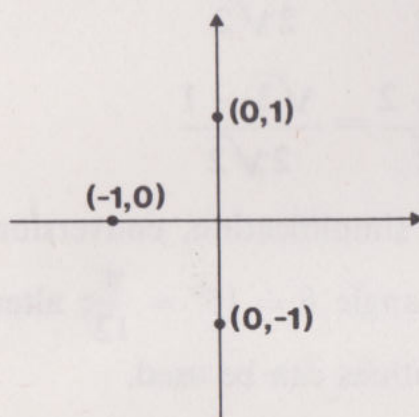
$$(x_1, y_1) = (-1, 0)$$

and

$$(x_2, y_2) = (0, -1)$$

then

$$(x_1, y_1) \otimes (x_2, y_2) = (0, 1)$$



$$\text{Arg}(-1, 0) = \pi \text{ and } \text{Arg}(0, -1) = \frac{3\pi}{2}.$$

$$\text{So their sum is } \frac{5\pi}{2}, \text{ but } \text{Arg}(0, 1) = \frac{\pi}{2}.$$

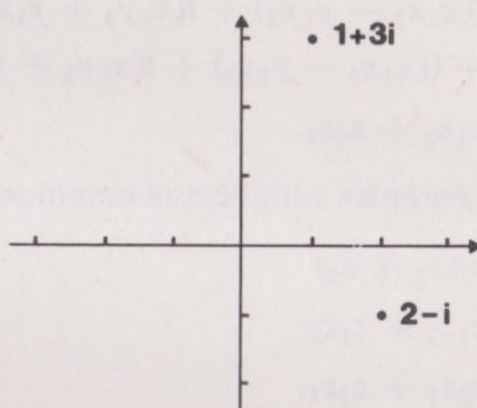
In fact, our previous experience (when we introduced the binary operation for the set  $A_1$ ) should tell us that  $\square$  is  $\oplus_{2\pi}$ , i.e. addition modulo  $2\pi$ .



## Section 9.4

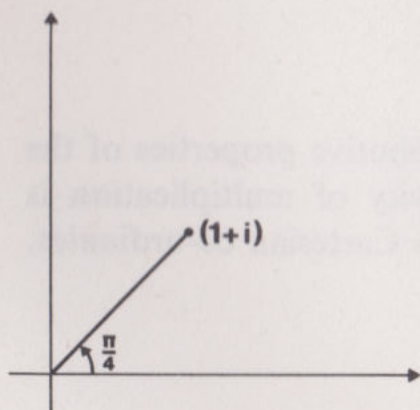
## Exercise 1

- (i) (a)  $5 + 5i$   
 (b)  $5$   
 (c)  $-11 + 13i$   
 (ii)  $-10 + 20i$   
 (iii)

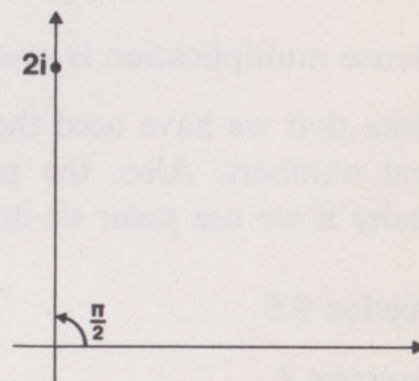


## Exercise 2

- (i)  $(1 + 2i)^2 = (1 + 2i)(1 + 2i) = 1 + 4i + 4i^2 = 1 + 4i - 4 = -3 + 4i$ .  
 Hence  $\text{Re}((1 + 2i)^2) = -3$ . Also  $\text{Im}((1 + 2i)^2) = 4$ . Notice particularly that the answer is 4, not  $4i$ .  
 (ii)  $1 + 2i + 3i^2 + i^3 = 1 + 2i - 3 - i = -2 + i$ . The real part is  $-2$ ; the imaginary part is 1.  
 (iii)



$$\text{Arg}(1 + i) = \frac{\pi}{4}.$$



$$\text{Arg}(1 + i)^2 = \frac{\pi}{2}.$$



*Exercise 3*

$$\begin{aligned}
z_1(z_2 + z_3) &= (x_1 + iy_1)((x_2 + iy_2) + (x_3 + iy_3)) \\
&= (x_1 + iy_1)((x_2 + x_3) + i(y_2 + y_3)) \\
&= (x_1(x_2 + x_3) - y_1(y_2 + y_3)) \\
&\quad + i(x_1(y_2 + y_3) + y_1(x_2 + x_3)) \\
&= (x_1x_2 + x_1x_3 - y_1y_2 - y_1y_3) \\
&\quad + i(x_1y_2 + x_1y_3 + y_1x_2 + y_1x_3) \\
&= ((x_1x_2 - y_1y_2) + i(x_1y_2 + y_1x_2)) \\
&\quad + ((x_1x_3 - y_1y_3) + i(x_1y_3 + y_1x_3)) \\
&= z_1z_2 + z_1z_3.
\end{aligned}$$

Since multiplication of complex numbers is commutative,

$$\begin{aligned}
(z_2 + z_3)z_1 &= z_1(z_2 + z_3) \\
&= z_1z_2 + z_1z_3 \\
&= z_2z_1 + z_3z_1.
\end{aligned}$$

Hence multiplication is distributive over addition. Note that we have used the associative and distributive properties of the real numbers.

*Exercise 4*

$$\begin{aligned}
z_1(z_2z_3) &= r_1(\cos \theta_1 + i \sin \theta_1) \\
&\quad \times (r_2r_3(\cos(\theta_2 + \theta_3) + i \sin(\theta_2 + \theta_3))) \\
&= r_1r_2r_3(\cos(\theta_1 + \theta_2 + \theta_3) + i \sin(\theta_1 + \theta_2 + \theta_3)) \\
&= r_1r_2(\cos(\theta_1 + \theta_2) + i \sin(\theta_1 + \theta_2)) \\
&\quad \times r_3(\cos \theta_3 + i \sin \theta_3) \\
&= (z_1z_2)z_3.
\end{aligned}$$

Hence multiplication is associative.

Note that we have used the associative and distributive properties of the real numbers. Also, the proof of the associativity of multiplication is easier if we use polar co-ordinates as opposed to Cartesian co-ordinates.

**Section 9.5***Exercise 1*

$$\begin{aligned}
\text{(i)} \quad (1 + i)(\tfrac{1}{2} - \tfrac{1}{2}i) &= \tfrac{1}{2} - \tfrac{1}{2}i + \tfrac{1}{2}i - \tfrac{1}{2}i^2 \\
&= \tfrac{1}{2} + \tfrac{1}{2} = 1
\end{aligned}$$



$$\begin{aligned} \text{(ii)} \quad (3 + 4i)\left(\frac{3}{25} - \frac{4}{25}i\right) &= \frac{9}{25} - \frac{12}{25}i + \frac{12}{25}i - \frac{16}{25}i^2 \\ &= \frac{9}{25} + \frac{16}{25} = 1 \end{aligned}$$

$$\begin{aligned} \text{(iii)} \quad (a + ib)(a - ib) &= a^2 - aib + iba - i^2b^2 \\ &= a^2 + b^2. \end{aligned}$$

(You should make sure that you understand which properties of  $C$  are being used at each step in the above manipulations.)

### Exercise 2

If you answered  $\theta$ , then you have forgotten that  $\arg$  is a one-many map. The correct answer is

$$\{\theta + 2k\pi, k \in \mathbb{Z}\}$$

## Section 9.6

### Exercise 1

We are given that

$$(a + ib)(x + iy) = 1,$$

and therefore

$$(ax - by) + i(bx + ay) = 1,$$

so that

$$ax - by = 1$$

and

$$bx + ay = 0.$$

Hence (multiplying the first equation by  $a$  and the second by  $b$ , and adding)

$$(a^2 + b^2)x = a$$

so that

$$x = \frac{a}{a^2 + b^2},$$

provided that  $a^2 + b^2 \neq 0$ .

In other words,  $a \neq 0$  and  $b \neq 0$ , so that we cannot allow  $a + ib = 0$ .

Similarly,  $y = \frac{-b}{a^2 + b^2}$  provided that  $a + ib \neq 0$ .



So the only number for which  $x + iy$  does not exist is 0, and this corresponds exactly to our experience with real numbers.

### Exercise 2

$$(i) \quad z\bar{z} = (x + iy)(x - iy) = x^2 - ixy + iyx - i^2y^2 = x^2 + y^2 = |z|^2$$

$$(ii) \quad \frac{1}{z} = \frac{x - iy}{x^2 + y^2} = \frac{\bar{z}}{|z|^2}$$

(iii) follows from (ii) and the definition of  $\div$ .

$$(iv) \quad z + \bar{z} = (x + iy) + (x - iy) = 2x = 2 \operatorname{Re} z$$

$$(v) \quad z - \bar{z} = (x + iy) - (x - iy) = 2iy = 2i \operatorname{Im} z.$$

### Exercise 3

We can use the formula

$$\frac{z_1}{z_2} = \frac{z_1\bar{z}_2}{|z_2|^2} \quad \text{or} \quad \frac{z_1\bar{z}_2}{z_2\bar{z}_2}.$$

Thus

$$(i) \quad \frac{1 - i}{3 + i} = \frac{(1 - i)(3 - i)}{3^2 + 1^2} = \frac{2 - 4i}{10} = \frac{1}{5} - \frac{2}{5}i$$

$$(ii) \quad \frac{1}{1 + i} + \frac{1 + i}{i} = \frac{1 - i}{1^2 + 1^2} + \frac{(-i)(1 + i)}{1^2} \\ = \frac{(1 - i) + (2 - 2i)}{2} = \frac{3}{2} - \frac{3}{2}i$$

### Exercise 4

(i)  $z_1 = 1$  and  $z_2 = -1$ , for example, show that this statement is false.

(ii) This statement is true. In polar co-ordinates we have

$$(r_1, \theta_1) \circ (r_2, \theta_2) = (r_1 r_2, \theta_1 + \theta_2),$$


 $z_1$ 

 $z_2$ 

 $z_1 \otimes z_2$ 

so that  $|z_1 \otimes z_2| = r_1 r_2 = |z_1| \times |z_2|$ .

*Exercise 5*

If  $z_1 = x_1 + iy_1$  and  $z_2 = x_2 + iy_2$ ,

then

$$z_1 + z_2 = (x_1 + x_2) + i(y_1 + y_2)$$

and

$$\begin{aligned}\overline{z_1 + z_2} &= (x_1 + x_2) - i(y_1 + y_2) \\ &= (x_1 - iy_1) + (x_2 - iy_2) \\ &= \bar{z}_1 + \bar{z}_2.\end{aligned}$$

So the first statement is true.

Similarly, it can be shown that

$$\overline{z_1 z_2} = \bar{z}_1 \bar{z}_2$$

i.e. that the second statement is true.

*Exercise 6*

Did you get stuck? If so, it's probably because you tried to work out  $z^n$  before taking the modulus.

Use the fact that  $|z^n| = |z|^n$ . The answer is  $(\frac{1}{2})^n$ .

**Section 9.7***Exercise 1*

With  $n = 3$ , we have

$$(\cos \theta + i \sin \theta)^3 = \cos 3\theta + i \sin 3\theta.$$

On the other hand, multiplying out, we have

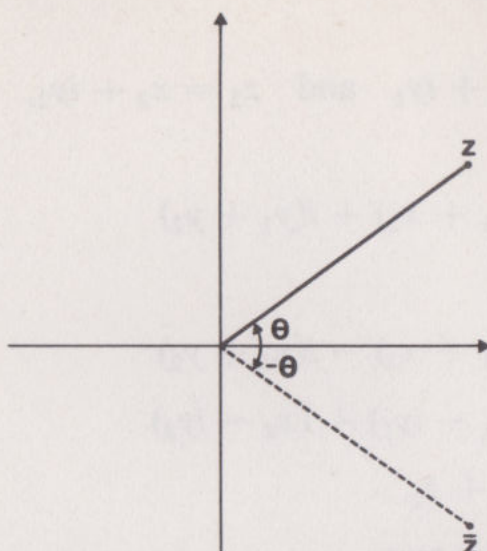
$$\begin{aligned}(\cos \theta + i \sin \theta)^3 &= \cos^3 \theta + 3i \cos^2 \theta \sin \theta \\ &\quad - 3 \cos \theta \sin^2 \theta - i \sin^3 \theta\end{aligned}$$

Equating the real parts of the two right-hand sides gives the required result.

*Exercise 2*

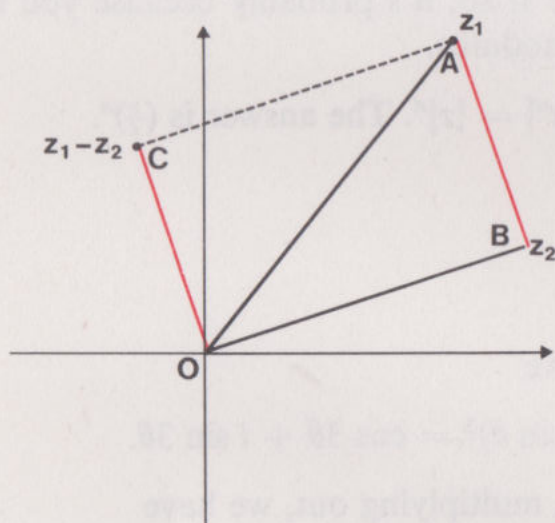
- (i) The point representing  $\bar{z}$  is the reflection in the real axis of the point representing  $z$ .





Note that if  $z$  has polar co-ordinates  $(r, \theta)$ , then  $\bar{z}$  has polar co-ordinates  $(r, -\theta)$ .

- (ii) If we represent  $z_1$  and  $z_2$  by geometric vectors, then the geometric vector representing  $z_1 + z_2$  is obtained by adding the geometric vectors. Therefore,  $z_1 - z_2$  is obtained by reversing the geometric vector representing  $z_2$  and then adding it to the one representing  $z_1$ .



In the diagram  $\underline{OA}$  represents  $z_1$  and  $\underline{OB}$  represents  $z_2$ . So that  $\underline{AC}$  represents  $-z_2$ . Hence  $z_1 - z_2$  is represented by  $\underline{OC} = \underline{BA}$  and  $|z_1 - z_2| = OC = BA$ , i.e.  $|z_1 - z_2|$  is the distance between the point representing  $z_1$  and the point representing  $z_2$ .

### Exercise 3

- (i) If  $z = x + iy$ , then

$$|z|^2 = x^2 + y^2,$$

and since  $y^2 \geq 0$

$$|z|^2 \geq x^2,$$

whence  $|z| \geq x = \operatorname{Re}(z)$ .

(ii) The numbers in brackets refer to the Summary on page 356.

$$\begin{aligned} |z_1 + z_2|^2 &= (z_1 + z_2)\overline{(z_1 + z_2)} && \text{(by (iv))} \\ &= (z_1 + z_2)(\bar{z}_1 + \bar{z}_2) && \text{(by (ii))} \\ &= z_1\bar{z}_1 + z_2\bar{z}_1 + z_1\bar{z}_2 + z_2\bar{z}_2 && \text{(by distributivity)} \\ &= |z_1|^2 + z_2\bar{z}_1 + z_1\bar{z}_2 + |z_2|^2 && \text{(by (iv)).} \end{aligned}$$

Notice that  $\overline{z_2\bar{z}_1} = \bar{z}_2z_1$  (by (iii))  $= \bar{z}_2z_1$  (by (vii)), i.e.  $z_2\bar{z}_1$  is the conjugate of  $z_1\bar{z}_2$ . And if we add a complex number to its conjugate, then we get simply twice its real part (by (v)). Hence

$$|z_1 + z_2|^2 = |z_1|^2 + 2 \operatorname{Re}(z_2\bar{z}_1) + |z_2|^2.$$

Applying the result to the complex number  $z_2\bar{z}_1$ , we have

$$\operatorname{Re}(z_2\bar{z}_1) \leq |z_2\bar{z}_1|$$

and hence

$$\begin{aligned} |z_1 + z_2|^2 &\leq |z_1|^2 + 2|z_2\bar{z}_1| + |z_2|^2 \\ &= |z_1|^2 + 2|z_2||\bar{z}_1| + |z_2|^2 && \text{(by (i))} \\ &= |z_1|^2 + 2|z_2||z_1| + |z_2|^2 && \text{(by (ix))} \\ &= (|z_1| + |z_2|)^2. \end{aligned}$$

Hence the modulus is positive or zero, we can now deduce that

$$|z_1 + z_2| \leq |z_1| + |z_2|.$$

Notice that this is an inequality between *real numbers*.

$$\begin{aligned} \text{(iii) } |z_1 - z_2|^2 &= (z_1 - z_2)\overline{(z_1 - z_2)} && \text{(by (iv))} \\ &= (z_1 - z_2)(\bar{z}_1 - \bar{z}_2) && \text{(by (ii))} \\ &= z_1\bar{z}_1 - z_2\bar{z}_1 - z_1\bar{z}_2 + z_2\bar{z}_2 && \text{(by distributivity)} \\ &= |z_1|^2 - 2 \operatorname{Re}(z_2\bar{z}_1) + |z_2|^2 && \text{(by (v))} \\ &\geq |z_1|^2 - 2|z_2\bar{z}_1| + |z_2|^2 && \text{(by (x))} \\ &= |z_1|^2 - 2|z_1||z_2| + |z_2|^2 && \text{(by (i) and (ix))} \\ &= (|z_1| - |z_2|)^2. \end{aligned}$$

Hence

$$|z_1 - z_2| \geq ||z_1| - |z_2||.$$



*Exercise 4*

The fact that the equation in  $R$  has an empty solution set can be deduced in many ways. To show that the given elements satisfy the equation in  $C$  is a matter of calculation. For instance, we simply substitute  $-1 + i\sqrt{3}$  into the left-hand side to obtain

$$(-1 + i\sqrt{3})^2 + 2(-1 + i\sqrt{3}) + 4$$

and this simplifies to zero.

We really also need to show that there are no other elements in the solution set, but we leave this.



## CHAPTER 10 COMPLEX FUNCTIONS

### 10.0 Introduction

In this chapter we shall take up the story of complex numbers from where we left it in Chapter 9. In that chapter we were mainly concerned with building an algebraic structure, and we did this by introducing a further binary operation on the vector space of ordered pairs of real numbers. We called this binary operation “multiplication”; in terms of number pairs it was defined by

$$(x_1, y_1) \otimes (x_2, y_2) = (x_1x_2 - y_1y_2, y_1x_2 + x_1y_2).$$

We introduced the notation  $(x, y) = x + iy$ , which is very useful because it enables us to work with this apparently complicated rule for multiplication without having to remember the above formula. With this notation we simply use the familiar rules of addition and multiplication, as in the algebra of real numbers: whenever we see  $i^2$  we replace it by  $-1$ , so that

$$\begin{aligned}(x_1 + iy_1)(x_2 + iy_2) &= x_1x_2 + i^2y_1y_2 + iy_1x_2 + ix_1y_2 \\ &= (x_1x_2 - y_1y_2) + i(y_1x_2 + x_1y_2).\end{aligned}$$

We call the set of all elements  $x + iy$  the set of *complex numbers*, and we denote this set by  $C$ . We saw in Chapter 9 that there is a subset of  $C$  which is isomorphic to  $R$  (for addition and multiplication) under addition and multiplication. So, by a small abuse of language, we can regard  $R$  as a subset of  $C$ . (Alternatively, we can say that  $C$  contains a “copy” of  $R$ .)

In sections 10.2–7 of this chapter we concentrate on functions from  $C$  to  $C$ , often called *complex functions*, and, in particular, on how they can be represented pictorially. It is possible to define complex forms of the well-known elementary functions  $\exp$ ,  $\sin$ ,  $\cos$ ,  $\tan$ ,  $\ln$ , etc., which reduce to their real forms when their domain is restricted to  $R$ , regarded as a subset of  $C$ , but in this chapter we will concentrate only on the function  $\exp$ . In these sections all we attempt to do is to give an introduction to some particular functions in order to give some “feel” for complex functions and their representation. In the final section we define and discuss the square roots and  $n$ th roots of a complex number.

### 10.1 Sets of Points in the Complex Plane

Often it is useful to specify a particular subset of the complex plane, and sometimes this can be done concisely in terms of argument, conjugate and modulus.



*Example 1*

What is the set  $\{z: z = \bar{z}\}$  on an Argand diagram?

If  $z = x + iy$  and  $z = \bar{z}$ , then

$$x + iy = x - iy$$

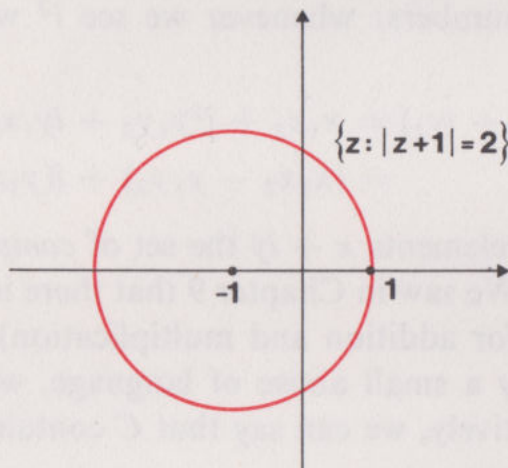
Equating imaginary parts, we get  $y = -y$ , so that  $y = 0$ .

In other words, the set coincides with the real axis.

*Example 2*

Indicate the set  $\{z: |z + 1| = 2\}$  on an Argand diagram.

If  $|z + 1| = 2$  then “the distance of  $z$  from  $-1$  is 2”, so that  $z$  lies on a circle with centre  $(-1, 0)$  and radius 2.

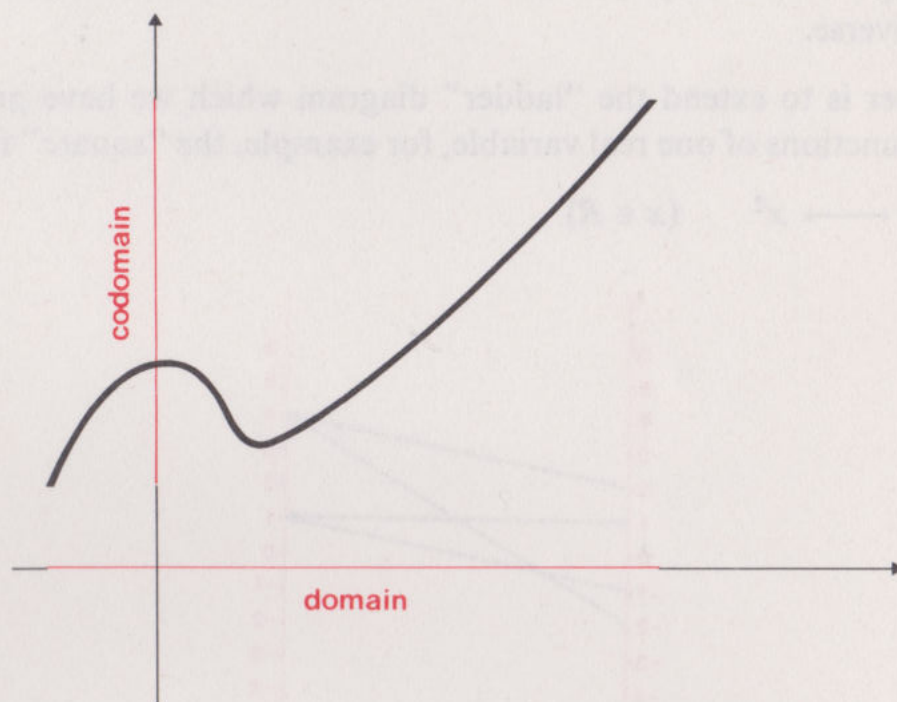
*Exercise 1*

Indicate the following sets on an Argand diagram.

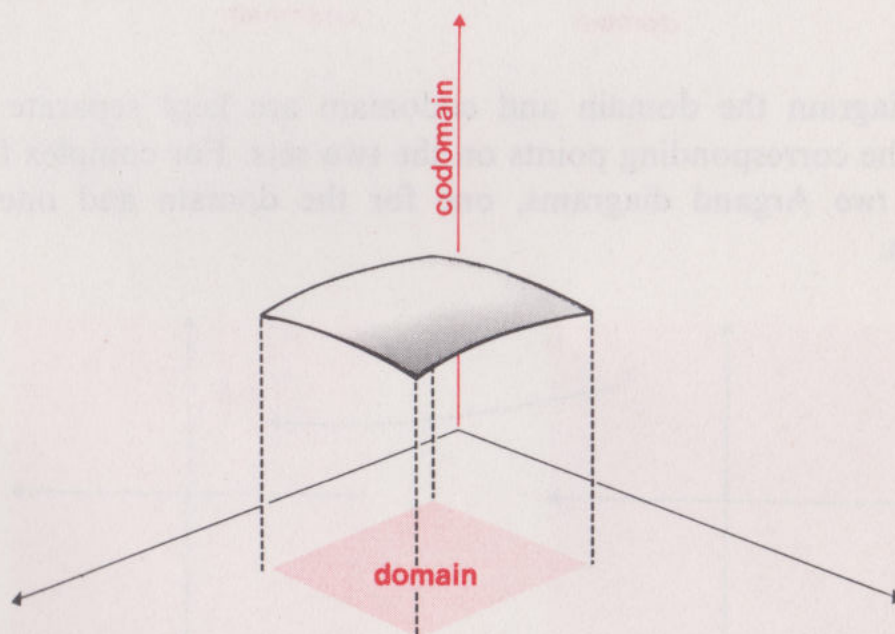
- (i)  $\{z: |z - 1| \leq 2\}$
- (ii)  $\left\{z: \operatorname{Arg} z = \frac{\pi}{4}\right\}$
- (iii)  $\{z: z + 2\bar{z} = 1\}$
- (iv)  $\left\{z: \operatorname{Arg}(z - 1) = \frac{\pi}{4}\right\}$
- (v)  $\{z: 0 \leq \operatorname{Arg} z \leq \pi\}$
- (vi)  $\left\{z: 0 \leq \operatorname{Arg} z \leq \frac{\pi}{2}, |z| \leq 1\right\}$
- (vii)  $\{z: |z - 1| = |z + 1|\}$
- (viii)  $\{z: |z - 1| \leq |z + 1|\}$

When looking at a particular real function in the past, one of the first things we did was to draw its graph. We shall now investigate what replaces the graph when we are dealing with complex functions.

We are very familiar with the fact that it is often possible to represent a function of one real variable by a graph:



We have also seen (Volume 2, Chapter 2) how it is often possible to represent a function of two real variables by a surface:



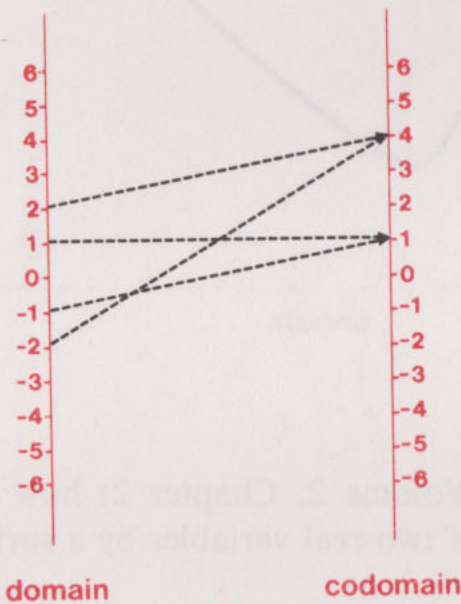
Representations of this kind are very useful because they give us an intuitive insight into the behaviour of the functions. We would like to



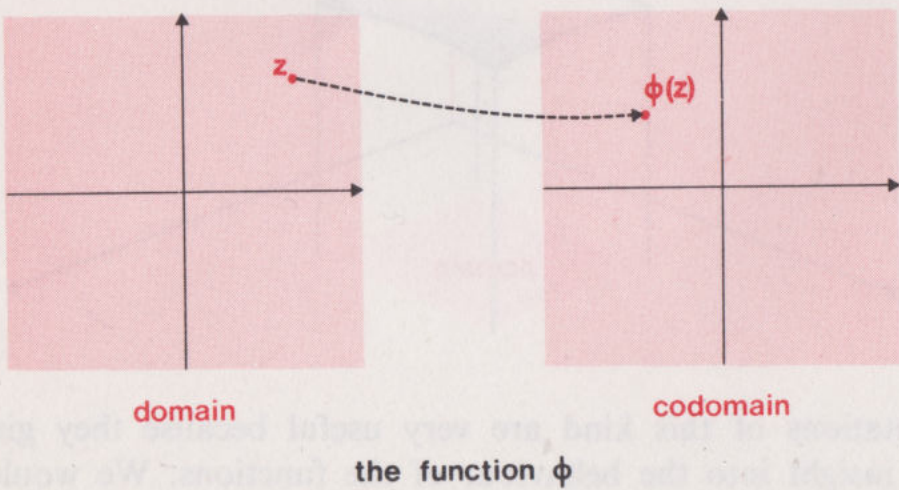
produce something similar for complex functions, but there is one major problem. Suppose that we have a function  $\phi: C \longrightarrow C$ . We know that  $C$  can be represented by an Argand diagram, but to obtain a representation of  $\phi$  similar to the graph of a real function, we need *two* copies of  $C$ , one for the domain of  $\phi$  and the other for the codomain. A direct extension of the graph would only be feasible for the inhabitants of a four-dimensional universe.

The answer is to extend the “ladder” diagram which we have previously used for functions of one real variable, for example, the “square” function:

$$x \longmapsto x^2 \qquad (x \in R).$$



In this diagram the domain and codomain are kept separate and we indicate the corresponding points on the two sets. For complex functions we have *two* Argand diagrams, one for the domain and one for the codomain.





Each point in the domain will be mapped to a corresponding point in the codomain by the complex function  $\phi$ , and we can indicate the corresponding points on the two Argand diagrams.

You may find this way of representing complex functions a little difficult at first. But once you have drawn a few such diagrams and used them to analyse simple functions, you will begin to find this method of representation very useful. In the next section we consider a simple example: our friend the “square” function.

## 10.2 The “Square” Function

As an example of the development of the representation of complex functions, we examine points and their images under the “square” function:

$$z \longmapsto z^2 \quad (z \in \mathbb{C}).$$

It is very often useful to let  $z = x + iy$  denote the (complex) variable in the domain, and  $w = u + iv$  denote the corresponding variable in the codomain, so that in this case we have

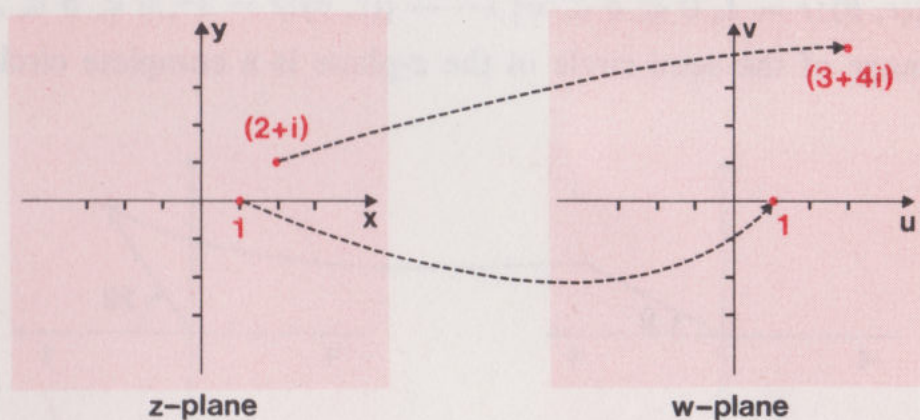
$$w = z^2 \quad (z \in \mathbb{C}).$$

The **domain** and **codomain** are then usually referred to as the ***z*-plane** and the ***w*-plane** respectively.

We get some intuitive feeling for the “square” function if we plot various points and their images;

$$\text{if } z = 1 \quad \text{then } w = 1,$$

$$\text{if } z = 2 + i \quad \text{then } w = 3 + 4i.$$



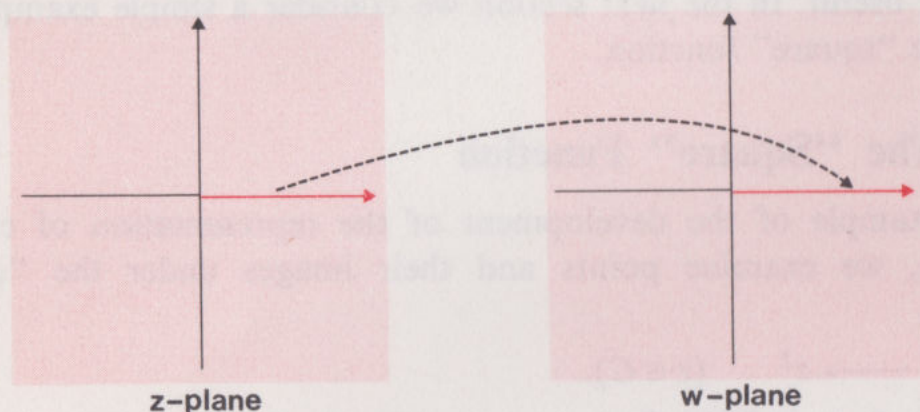
Isolated points are not sufficient to convey the behaviour of the function, but if we consider the images of a *set of points* then we can often see the properties of the function more clearly. For example, what is the image of the set of points which lie on the positive real axis in the *z*-plane?



The positive real axis in the  $z$ -plane is the set  $\{(x, 0): x > 0\}$ , and since  $w = z^2$  we know that in the image set

$$w = u + iv = (x + i0)^2 = x^2.$$

Therefore the image set is the set of points for which  $u$  is positive and  $v = 0$ , i.e.  $\{(u, 0): u > 0\}$ , the positive real axis in the  $w$ -plane.



Now let us examine the image of a semi-circle in the upper half-plane with centre at the origin and radius 1. This set is most easily specified in terms of polar co-ordinates;\* it is the set of complex numbers whose polar co-ordinates belong to the set

$$\{(r, \theta): r = 1, 0 \leq \theta \leq \pi\}.$$

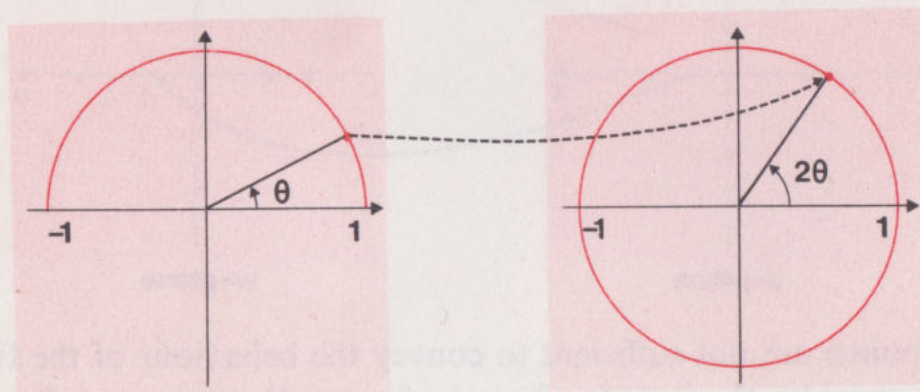
Since the “square” function is also very easily specified in terms of polar co-ordinates:

$$(r, \theta) \longmapsto (r^2, 2\theta),$$

it seems best to work entirely in this co-ordinate system. It follows that

$$\{(r, \theta): r = 1, 0 \leq \theta \leq \pi\} \longmapsto \{(r, \theta): r = 1^2, 0 \leq \theta \leq 2\pi\},$$

so the image of the semi-circle in the  $z$ -plane is a complete circle in the  $w$ -plane.



\* We use black brackets for polar co-ordinates now, as is usual in the mathematical literature. It should be clear from the context which system of co-ordinates we are using.

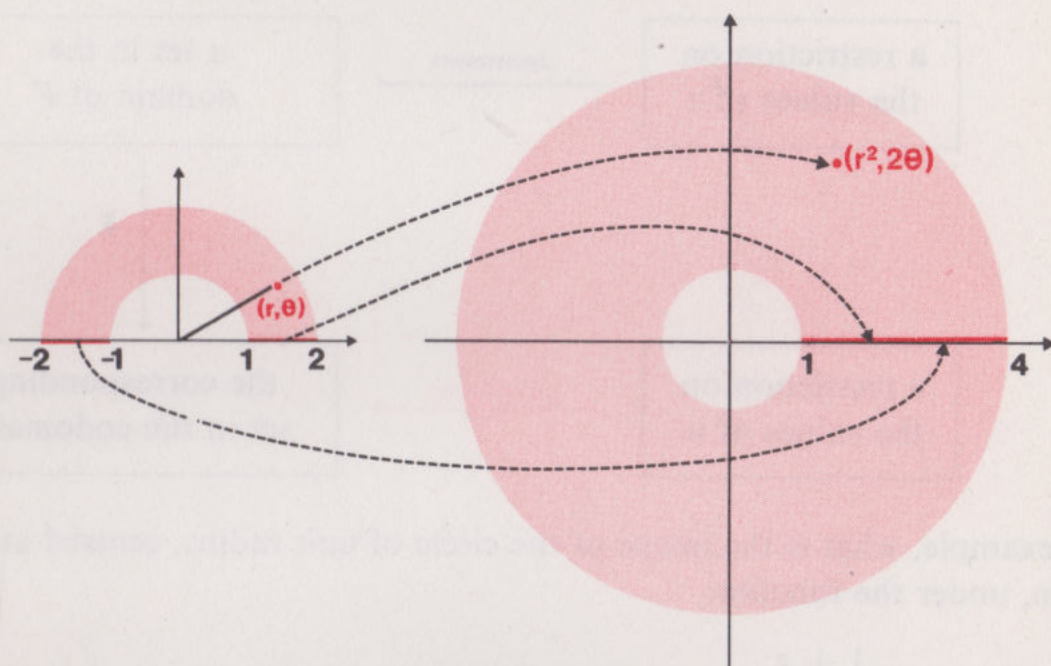


We can also obtain useful information from looking at the images of *regions* in the domain. For example, consider the set of complex numbers specified by the set

$$\{(r, \theta): 1 \leq r \leq 2, 0 \leq \theta \leq \pi\},$$

which represents an annular region. Each point specified by  $(r, \theta)$  in the domain is mapped to  $(r^2, 2\theta)$  by the “square” function, so the image set is the set of complex numbers

$$\{(r, \theta): 1 \leq r \leq 4, 0 \leq \theta \leq 2\pi\}.$$



This set of complex numbers is more neatly expressed as the set

$$\{z: 1 \leq |z| \leq 4\}.$$

### Exercise 1

Draw the image of the line segment

$$\{z: 1 \leq y \leq 2, x = 0\}$$

under the “square” function.

## 10.3 Representation of Complex Functions

You may find it difficult to see how we actually determine the image set under a particular function. Sometimes it is easy geometrically with little or no algebraic manipulation, as in the case of the “square” function,

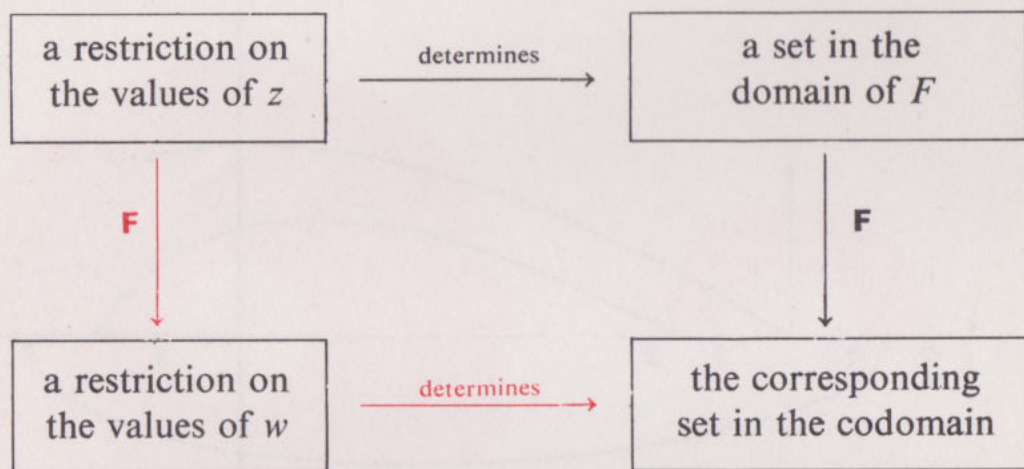


but often we are forced to use an algebraic approach, interpreting the algebra at the end to obtain the geometric picture.

Suppose that we have a function

$$F: z \longmapsto w \quad (z \in C).$$

A set in the domain is determined by some restriction on the values of  $z$ ; under the function  $F$  this is converted to a restriction on the values of  $w$ , and hence we determine our set in the codomain. Diagrammatically we have:



For example, what is the image of the circle of unit radius, centred at the origin, under the function

$$z \longmapsto \frac{1+z}{1-z} \quad (z \in C, z \neq 1)?$$

The circle is determined by the equation

$$|z| = 1 \quad z \neq 1,$$

and this is our restriction on  $z$ . (Notice that we cannot have the complete circle, because we have not included  $z = 1$  in the domain of our function.)

We put

$$w = \frac{1+z}{1-z}$$

so that

$$w - zw = 1 + z$$

and hence

$$z = \frac{w-1}{w+1} \quad w \neq -1.$$

We can now substitute for  $z$  in the equation representing the restriction, to obtain

$$\left| \frac{w-1}{w+1} \right| = 1 \quad w \neq -1,$$

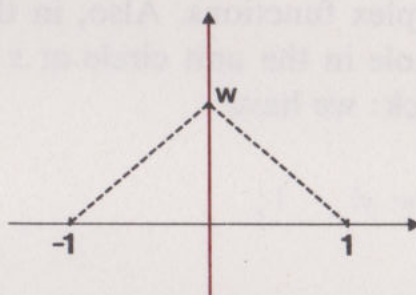
and this is a restriction on the values of  $w$  in the codomain. It only remains to give a geometric interpretation of this set. If we rearrange the equation, we get

$$|w-1| = |w+1| \quad w \neq -1.$$

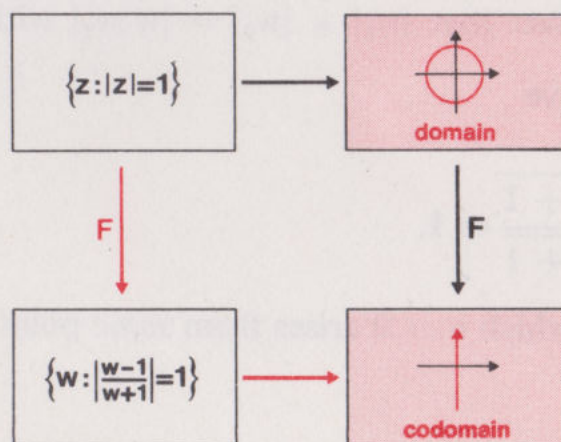
We know (from Exercise 9.7.2) that  $|w-a|$  means the distance of  $w$  from  $a$ . We can now read the equation as

“the distance of  $w$  from 1 equals the distance of  $w$  from  $-1$ , where  $w \neq -1$ ”.

The points equidistant from 1 and  $-1$  clearly lie on the perpendicular bisector of the line joining 1 and  $-1$ , which is the imaginary axis. (The condition  $w \neq -1$  does not exclude any of these points.)



Using the same diagrammatic representation as in the general discussion, we have



Notice that at one stage we need to find  $z$  in terms of  $w$ , which effectively means finding the reverse function. In our example the function is one-one,



but in other cases it might be less straightforward; for example, this technique might not be effective for the “square” function which is a many-one function.

There is one point which we have glossed over in the discussion so far, which is best explained in terms of the example we have just considered.

We know that every point of the unit circle (except the point  $z = 1$ ) in the  $z$ -plane maps to a point on the imaginary axis in the  $w$ -plane. We can express this by

$$|z| = 1, \quad z \neq 1 \text{ implies } |w - 1| = |w + 1|, \quad w \neq -1$$

$$\text{whence } u = 0,$$

where  $w = u + iv$ .

But we have not checked that the image of the unit circle is the *whole* of the imaginary axis, that is,

$$u = 0 \text{ implies } |z| = 1 \text{ and } z \neq 1.$$

Intuition would suggest that it is so, but intuition can be very deceptive when dealing with complex functions. Also, in this particular case there is rather an awkward hole in the unit circle at  $z = 1$ . As it happens, it is not very difficult to check: we have

$$z = \frac{w - 1}{w + 1} \quad w \neq -1;$$

putting  $u = 0$ , we get

$$z = \frac{iv - 1}{iv + 1}$$

whence, using the fact that  $|w_1| \times |w_2| = |w_1 w_2|$  with  $w_2 = iv - 1$  and  $w_2 = \frac{1}{iv + 1}$ , we have

$$|z| = \frac{\sqrt{v^2 + 1}}{\sqrt{v^2 + 1}} = 1.$$

So *every* point for which  $u = 0$  arises from *some* point for which  $|z| = 1$ .

### Exercise 1

Find the image of each of the following sets under the “square” function

$$z \longmapsto z^2 \quad (z \in \mathbb{C}).$$



- (i)  $\{z: x < 0, y \in \mathbb{R}\}$
- (ii)  $\{z: x = 1, y \in \mathbb{R}\}$
- (iii)  $\{z: x \in \mathbb{R}, y = 1\}$

### Exercise 2

Find the image of the circle centred at the origin with unit radius under the function

$$z \longmapsto 2z + 3 \quad (z \in \mathbb{C}).$$

### Invariance

We first mentioned the concept of invariance in section 5.4. Invariance is another of the notions which we introduce in this volume because it occurs widely in mathematics. Often in mathematics it is very interesting and useful to ask what is left undisturbed, or invariant, under a function. For example, the real exponential function  $x \longmapsto e^x$  ( $x \in \mathbb{R}$ ) is invariant under the differentiation operator  $D$ . We have in fact already seen many different examples of invariance.

In the context of complex functions, invariant points and sets of points can be of considerable assistance in the visualization of the functions. An **invariant point** of the complex function  $f$  is a point  $a$  such that

$$f(a) = a.$$

An **invariant set** under the function  $f$  is a set  $A$  such that

$$f(A) = A.$$

Notice that in the latter case the individual points of  $A$  need not themselves be invariant; that is, we do not require that

$$f(a) = a \quad \text{for all } a \in A,$$

but we do require that

$$\{f(a): a \in A\} = A.$$

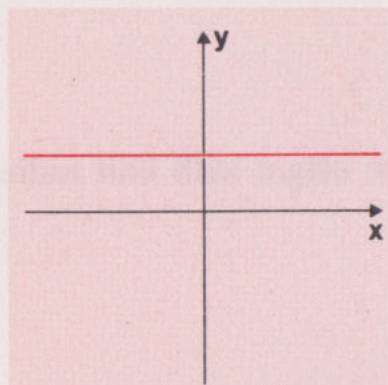
Notice also that, although we use two Argand diagrams to represent a complex function, the idea of invariance is expressed more naturally in terms of one; i.e. we are regarding the codomain as superimposed on the domain. For example, consider the function

$$z \longmapsto z + 1 \quad (z \in \mathbb{C}),$$

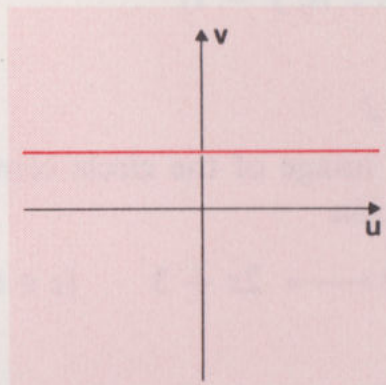
which translates every point  $z$  one unit parallel to the real axis. We know



that this function has no invariant points, but it does have invariant sets of points. For example, any line parallel to the real axis is invariant.



domain



codomain

Although each point moves to a new point, the set of all points lying on such a line remains unchanged.

## 10.4 The Exponential Function

### Introduction

In this section we consider the complex exponential function. In doing so we use many of the points discussed in section 10.3 and in Chapter 9.

The definition of multiplication in  $C$  has a natural interpretation in terms of the geometric notions of scaling and rotation. The polar co-ordinate system is very useful in this context because multiplication of complex numbers in polar form is particularly easy. To illustrate this statement we shall revise a result, discussed in Chapter 9, which we shall use again in this chapter. Suppose that we take a particular complex number  $z = x + iy$  corresponding to the polar co-ordinates  $(r, \theta)$ , so that

$$z = x + iy = r \cos \theta + ir \sin \theta;$$

then we know that, for any positive integer  $n$ ,  $z^n$  corresponds to the polar co-ordinates  $(r^n, n\theta)$ . In other words,

$$\begin{aligned} z^n &= r^n \cos n\theta + ir^n \sin n\theta \\ &= r^n (\cos n\theta + i \sin n\theta). \end{aligned}$$

But we know that

$$\begin{aligned} z^n &= (r \cos \theta + ir \sin \theta)^n \\ &= r^n (\cos \theta + i \sin \theta)^n \end{aligned}$$



and therefore

$$\cos n\theta + i \sin n\theta = (\cos \theta + i \sin \theta)^n.$$

This result is a special case of De Moivre's Theorem, and we shall use it in developing a definition of the complex exponential function.

### Extending the Domain

In Volume 1, Chapter 5 we defined the exponential function with domain  $R$ , and our object now is to extend the domain of the exponential function to  $C$ . You will also remember that  $\exp x$  could alternatively be written as  $e^x$  (where  $e \simeq 2.71828$ ).

Obviously the extended exponential function

$$z \longmapsto \exp z \quad (z \in C)$$

must coincide with the original function

$$x \longmapsto \exp x \quad (x \in R)$$

when the domain is restricted to  $R$ . But this in itself is not a sufficient guide to suggest a definition of  $\exp z$ , so we shall specify some of the properties of  $\exp x$  which we would like  $\exp z$  to have also.

Unfortunately, the most obvious approach is not possible. In Volume 1, Chapter 5 we defined the exponential function by

$$\exp: x \longmapsto \lim_{k \text{ large}} \left( 1 + \frac{x}{k} \right)^k \quad (x \in R \text{ and } k \in Z^+).$$

But we have not defined limits in the context of complex numbers (although we could), so we cannot define the complex exponential function to be

$$\exp: z \longmapsto \lim_{k \text{ large}} \left( 1 + \frac{z}{k} \right)^k \quad (z \in C \text{ and } k \in Z^+),$$

although, if we gave an appropriate meaning to the limit, this would be a satisfactory definition.

In Volume 1, Chapter 8, we found a characteristic property of the exponential function:

$$(\exp)' = \exp.$$

But unfortunately we have not defined the derived function of a complex function (although we could).



In Volume 2, Chapter 5 we found a convergent series for the exponential function:

$$\exp x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots \quad (x \in R)$$

But we have not defined convergence of an infinite complex series (although we could), so we can't just replace  $x$  by  $z$  and  $R$  by  $C$  to obtain a definition of the complex exponential function.

We have recounted this tale of woe because, had we the necessary ancillary ideas, each of these three possibilities for a definition could provide us with a more satisfactory approach than the one we actually adopt. But we can make up what we lack in expertise by bravado.

In Volume 1, Chapter 5 we discussed a very important property of the exponential function:

$$\exp(x_1 + x_2) = \exp x_1 \times \exp x_2 \quad (x_1, x_2 \in R),$$

that is, the real exponential function is a morphism (isomorphism) of  $(R, +)$  to  $(R^+, \times)$ . Let us agree to preserve this property for a start; that is, we require of the complex exponential function that

$$\exp(z_1 + z_2) = \exp z_1 \exp z_2 \quad (z_1, z_2 \in C).$$

This means that the complex exponential function is a morphism (perhaps not an isomorphism) of  $(C, +)$  to  $(C_1, \otimes)$ , where  $C_1$  is some, as yet undetermined, subset of  $C$  and  $\otimes$  is the symbol we introduced for complex multiplication in Chapter 9.

There are some immediate consequences. Let

$$z_1 = x \quad \text{and} \quad z_2 = iy \quad (x, y \in R);$$

then

$$\exp(x + iy) = \exp x \exp(iy),$$

so for any complex number  $z$ , we know that

$$\exp z = e^x \exp(iy).$$

We can see that  $\exp z$  falls into two parts, one of which is the real number  $e^x$ . Our problem now is to find a suitable definition of  $\exp(iy)$ , so we investigate this part further.

We have assumed that  $\exp z$  is a complex number, so we can write

$$\exp(iy) = f(y) + ig(y),$$

where  $f$  and  $g$  are real functions.



Now let  $n$  be any positive integer. Then

$$\exp(iny) = (\exp(iy))^n,$$

by repeated application of our morphism property. So that

$$f(ny) + ig(ny) = (f(y) + ig(y))^n.$$

Compare this with the special case of De Moivre's Theorem cited above:

$$\cos n\theta + i \sin n\theta = (\cos \theta + i \sin \theta)^n.$$

The suggestion is clear, but there is no "proof" for any conclusion. So we resort to bravado and define

$$\exp(iy) = \cos y + i \sin y \quad (y \in \mathbb{R}).$$

So we define the complex exponential function by

$$\exp: x + iy \longmapsto e^x (\cos y + i \sin y) \quad (x + iy \in \mathbb{C}).$$

The first thing we must do is to check that the complex exponential function does reduce to the real exponential function when the domain is restricted to  $\mathbb{R}$ . Indeed it does, for then  $y = 0$  and

$$\exp(x + i0) = e^x (\cos 0 + i \sin 0),$$

so that

$$\exp x = e^x.$$

It seems, therefore, that this function has some of the desirable properties which we could expect from an extension of the real exponential function. If the complex differential calculus had been defined we would also find that

$$(z \longmapsto \exp z)' = (z \longmapsto \exp z),$$

(where we use ' to indicate the derived function).

In order to achieve consistency with the real exponential function, it is quite common to put

$$\exp z = e^z.$$

(Strictly speaking, this does not mean " $e$  to the power  $z$ " because we have not defined what is meant by numbers raised to complex powers.)

The equation

$$\exp(iy) = \cos y + i \sin y$$



or

$$e^{iy} = \cos y + i \sin y$$

is known as **Euler's formula**.

There is a mathematical oddity which is interesting. If we put  $y = \pi$  in Euler's formula, then we have

$$e^{i\pi} = (\cos \pi + i \sin \pi) = -1,$$

so that

$$e^{i\pi} + 1 = 0.$$

This equation contains five of the most significant numbers in the history of mathematics in a neat and tidy formula:  $e$ ,  $i$ ,  $\pi$ , 1 and 0. We could devote a chapter to each of them.

It is important to notice that the complex exponential function embodies a neat relationship between a complex number  $z = x + iy$  and the corresponding polar co-ordinates  $(r, \theta)$ . We know already that

$$x = r \cos \theta,$$

and

$$y = r \sin \theta,$$

so that

$$z = r (\cos \theta + i \sin \theta);$$

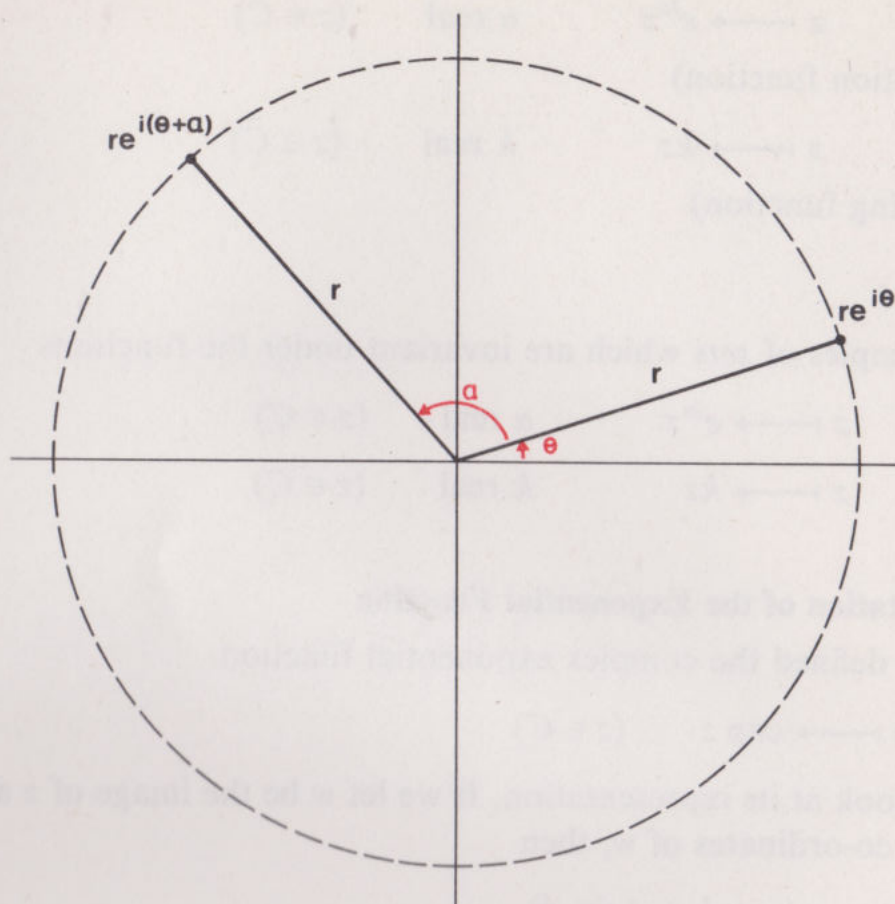
using Euler's formula we can now write this as

$$z = re^{i\theta}.$$

We shall use  $z$  as an abbreviation for each of the various ways of writing a complex number:  $(x, y)$ ,  $x + iy$ ,  $(r, \theta)$ ,  $r (\cos \theta + i \sin \theta)$  and  $re^{i\theta}$ . This will not lead to any confusion for each representation refers to the same point  $z$  in the complex plane.

Notice particularly that multiplying a complex number by  $e^{i\alpha}$  (for any real number  $\alpha$ ) has the effect of rotating the corresponding geometric vector anti-clockwise through an angle  $\alpha$  about the origin; in other words, increasing the argument of the complex number by  $\alpha$ , for

$$\begin{aligned} e^{i\alpha}z &= re^{i\alpha}e^{i\theta}, \\ &= re^{i(\alpha+\theta)}, \\ &= r (\cos (\alpha + \theta) + i \sin (\alpha + \theta)). \end{aligned}$$

*Exercise 1*

- (i) Show that if  $z = e^{i\theta}$ , then  $\bar{z} = e^{-i\theta}$ .
  - (ii) Show that  $|e^{i\theta}| = 1$  for all real numbers  $\theta$ .
  - (iii) Show that  $|e^z| = e^x$ .
  - (iv) Show that  $e^{z+i2\pi} = e^z$  for all complex numbers  $z$ .
  - (v) Find the set of all complex numbers  $z$  for which  $e^z = 1$ .
- (HINT: You may find it helpful to refer to the Summary on page 356.)

*Exercise 2*

Show that if

$$z = re^{i\theta} \quad r \neq 0,$$

then

$$\frac{1}{z} = \frac{1}{r}e^{-i\theta}.$$

*Exercise 3*

Which *points* are invariant under the following functions?

- (i)  $z \longmapsto z + z_0 \quad z_0 \in \mathbb{C} \quad (z \in \mathbb{C})$   
(translation function)



- (ii)  $z \longmapsto e^{i\alpha}z$        $\alpha$  real      ( $z \in C$ )  
 (rotation function)
- (iii)  $z \longmapsto kz$        $k$  real      ( $z \in C$ )  
 (scaling function)

**Exercise 4**

Find examples of *sets* which are invariant under the functions

- (i)  $z \longmapsto e^{i\alpha}z$        $\alpha$  real      ( $z \in C$ )
- (ii)  $z \longmapsto kz$        $k$  real      ( $z \in C$ )

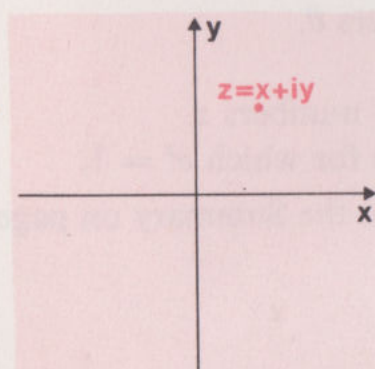
**Representation of the Exponential Function**

We have defined the complex exponential function

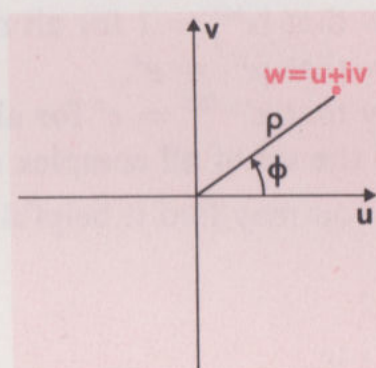
$$z \longmapsto \exp z \quad (z \in C)$$

we now look at its representation. If we let  $w$  be the image of  $z$  and  $(\rho, \phi)$  be polar co-ordinates of  $w$ , then

$$\begin{aligned} w &= \rho (\cos \phi + i \sin \phi) \\ &= \rho e^{i\phi} \end{aligned}$$



domain



codomain

But

$$w = e^z = e^x e^{iy}$$

so that

$$\rho = e^x$$

and

$$e^{i\phi} = e^{iy}$$

Consider now the image of a horizontal strip parallel to the  $x$ -axis in the  $z$ -plane, specified by

$$\{z: y \in [y_1, y_2], 0 \leq y_1 < y_2 < 2\pi\}.$$

If we choose the value of  $\phi$  to be  $\text{Arg } w$ , i.e.  $\phi \in [0, 2\pi[$ , then it follows from

$$e^{i\theta} = e^{iy}$$

that

$$\phi = y.$$

Since

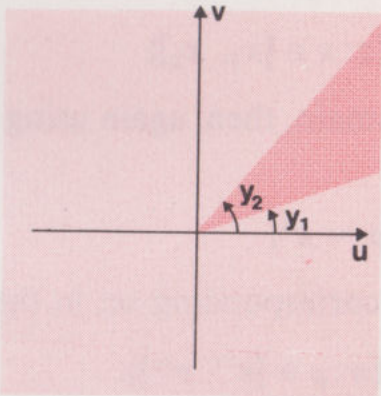
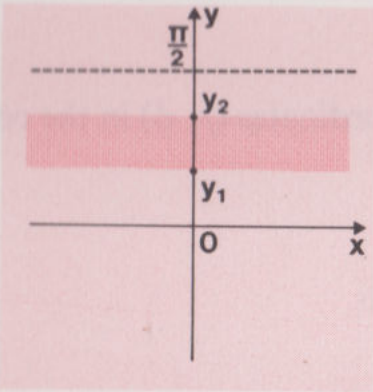
$$y \in [y_1, y_2],$$

we have

$$\phi \in [y_1, y_2],$$

which is a restriction on the values of  $w$  corresponding to the horizontal strip in the  $z$ -plane.

The set  $\{w: \phi \in [y_1, y_2]\}$  is a wedge shape with its apex at the origin and apex angle  $y_2 - y_1$ .

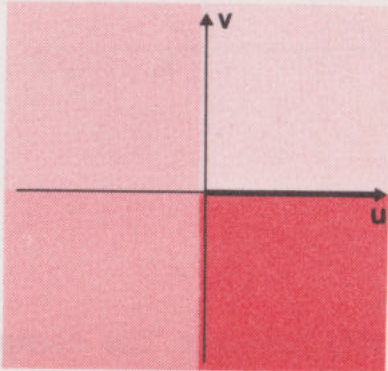
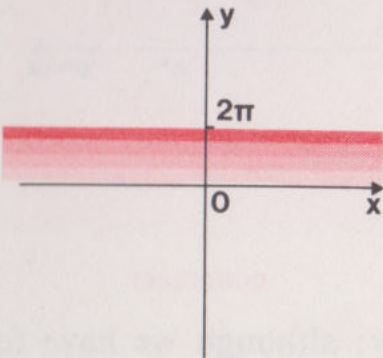


The exponential function  
 $z \mapsto e^z$

domain

codomain

If we allow the horizontal strip to be of width  $2\pi$  (excluding  $2\pi$  but including 0) then it will map to the complete  $w$ -plane.



domain

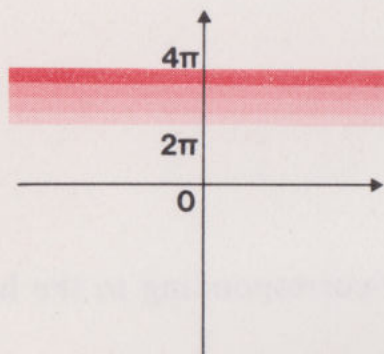
codomain



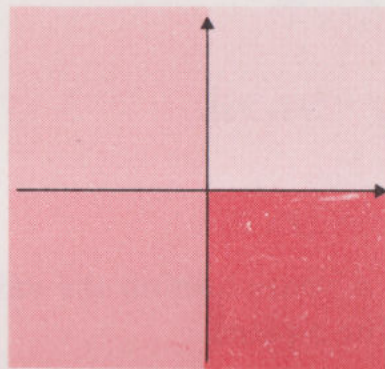
If we map the horizontal strip

$$\{z: y \in [2\pi, 4\pi[ \}$$

(also of width  $2\pi$ ) then the image is again the complete  $w$ -plane.



domain



codomain

It is also interesting to see what happens to a strip parallel to the  $y$ -axis under the exponential function. Suppose that we take the set

$$\{z: x \in [x_1, x_2]\}$$

in the domain, then, again using polar co-ordinates  $(\rho, \phi)$  in the codomain, we have

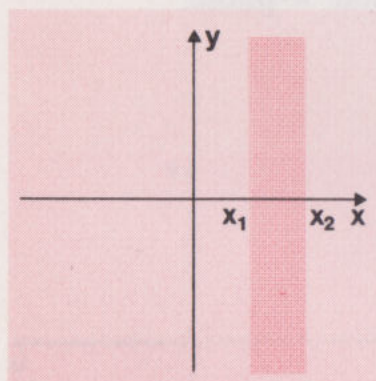
$$\rho = e^x,$$

and the corresponding set in the  $w$ -plane is

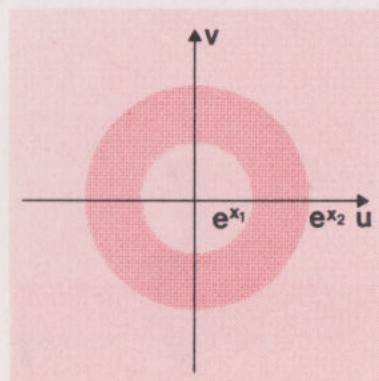
$$\{w: \rho \in [e^{x_1}, e^{x_2}]\}.$$

(Notice that we are using the fact that if  $x_2 > x_1$  then  $e^{x_2} > e^{x_1}$ .)

Since  $\rho = |w|$ , we see that the image is the annulus lying between the circles centred at the origin, with radii  $e^{x_1}$  and  $e^{x_2}$ .



domain

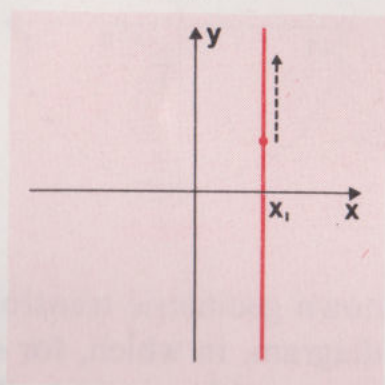


codomain

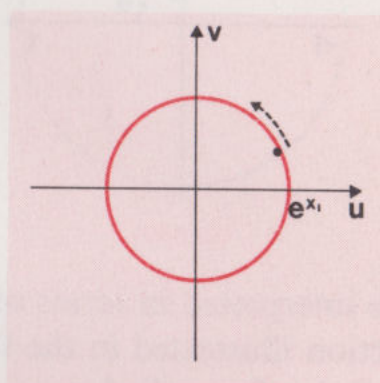
This is not, however, the complete story: although we have found the

image of the set  $\{z: x \in [x_1, x_2]\}$ , there is one other important consideration. Suppose that a point on the line  $x = x_1$  starts on the  $x$ -axis and moves up the line (i.e. in the direction of the positive  $y$ -axis). What is the locus of the image point in the  $w$ -plane, and how do the corresponding points move?

We can easily see from our previous discussion that the locus in the  $w$ -plane is simply the circle  $\rho = |w| = e^{x_1}$ ; but how do the corresponding points move around this circle?



domain



codomain

As  $y$  increases from 0 to  $2\pi$ , so  $\phi = \text{Arg } w = y$  increases from 0 to  $2\pi$ , and the image point travels all the way round the circle in the  $w$ -plane. In fact, the image point  $w$  makes a complete circuit of the circle for each interval of length  $2\pi$  on the line.

The function  $z \mapsto \exp z$  is many-one with a vengeance; an infinite number of points in the domain map to each point of the codomain.

### Exercise 5

- (i) Find the set of elements which map to  $e^{x_1}$ ,  $x_1 \in \mathbb{R}$ , under the function

$$z \mapsto \exp z.$$

- (ii) Find the set of elements which map to  $e^{x_1}e^{i\theta}$  under the function

$$z \mapsto \exp z.$$

## 10.5 The Function $z \mapsto \frac{1}{z}$

We shall denote the function

$$z \mapsto \frac{1}{z} \quad (z \in \mathbb{C}, z \neq 0)$$

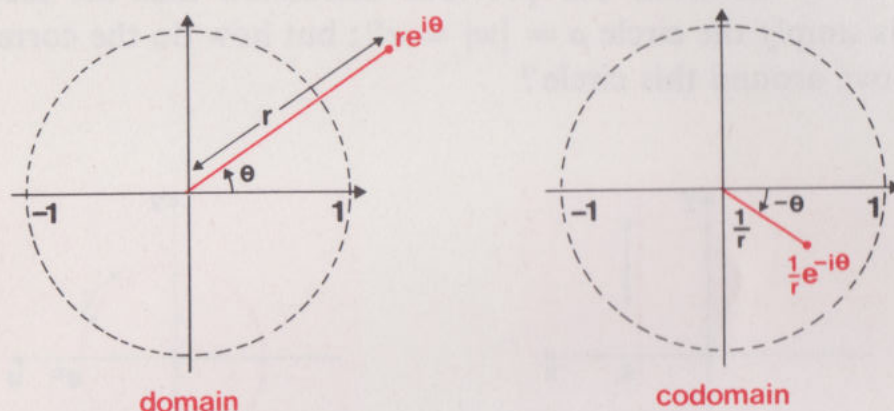
by  $\Phi$ .



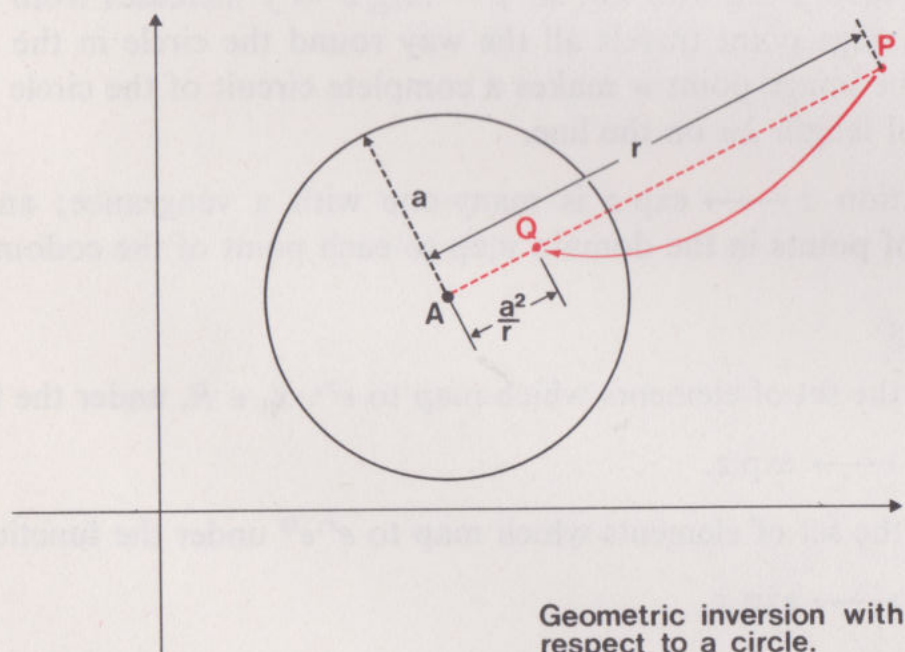
If we use the form  $z = re^{i\theta}$ , for convenience, then

$$\Phi: re^{i\theta} \mapsto \frac{1}{re^{i\theta}} = \frac{1}{r}e^{-i\theta}.$$

(See Exercise 10.4.2.)



$\Phi$  can be interpreted in terms of a well-known geometric transformation. The function illustrated in the following diagram, in which, for example,  $P \mapsto Q$ , is often called a **geometric inversion** with respect to the circle, centre  $A$ .



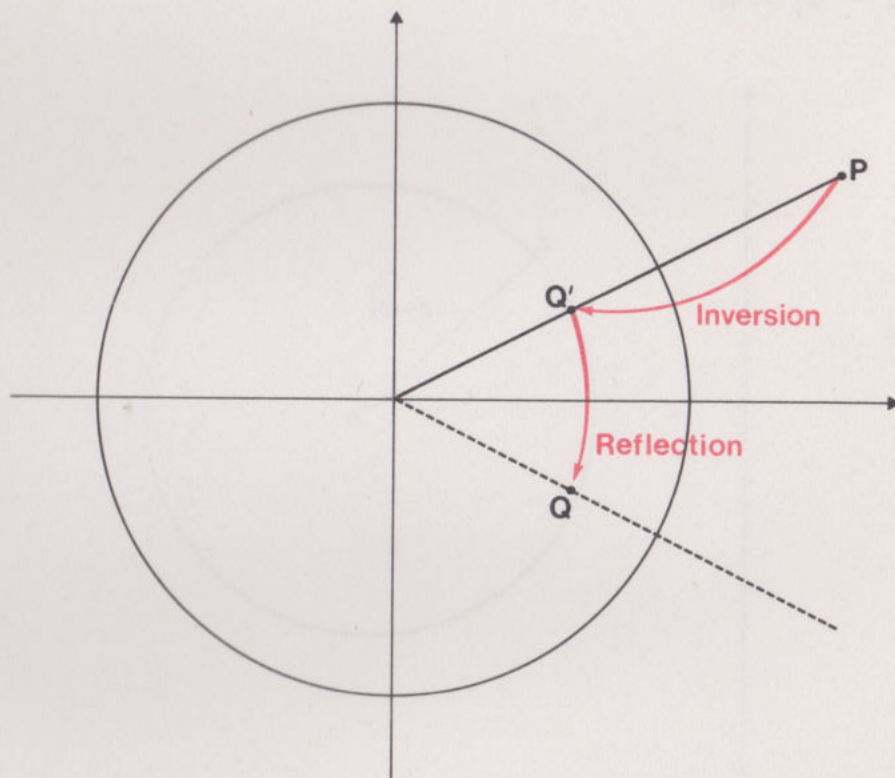
$P$  and  $Q$  are known as a **pair of inverse points** with respect to the circle. These points have the property that  $A$ ,  $P$  and  $Q$  are collinear and  $AQ \cdot AP = a^2$ , where  $a$  is the radius of the circle. It is clear that the function is one-one, and that points inside the circle map to points outside the circle and vice versa, except for the centre of the circle which has no image since it is not included in the domain of the function. Points on

the circle map to themselves, that is, they are invariant points of the transformation.

If we now look at our function  $\Phi$ , then we see that it is very similar to a geometric inversion. If  $P$  represents  $z$ ,  $Q$  represents  $\frac{1}{z}$ ,  $A$  is the origin  $O$  and  $a = 1$ , then

$$OP \times OQ = |z| \times \left| \frac{1}{z} \right| = 1,$$

which suggests inversion in the unit circle centred at the origin. The only trouble is that  $O$ ,  $P$  and  $Q$  are not, in general, collinear, since  $\text{Arg} \left( \frac{1}{z} \right) = -\text{Arg } z$ . If  $P$  and  $Q'$  are inverse points with respect to the circle  $|z| = 1$ , then  $Q$  is the reflection of  $Q'$  in the real axis.



Under the function  $\Phi: z \mapsto \frac{1}{z}$  we have, for example,  $P \mapsto Q$  in the last diagram. So we see that

the function  $z \mapsto \frac{1}{z}$  can be interpreted as a geometric inversion with respect to the circle centred at the origin, with unit radius, followed by a reflection in the  $x$ -axis



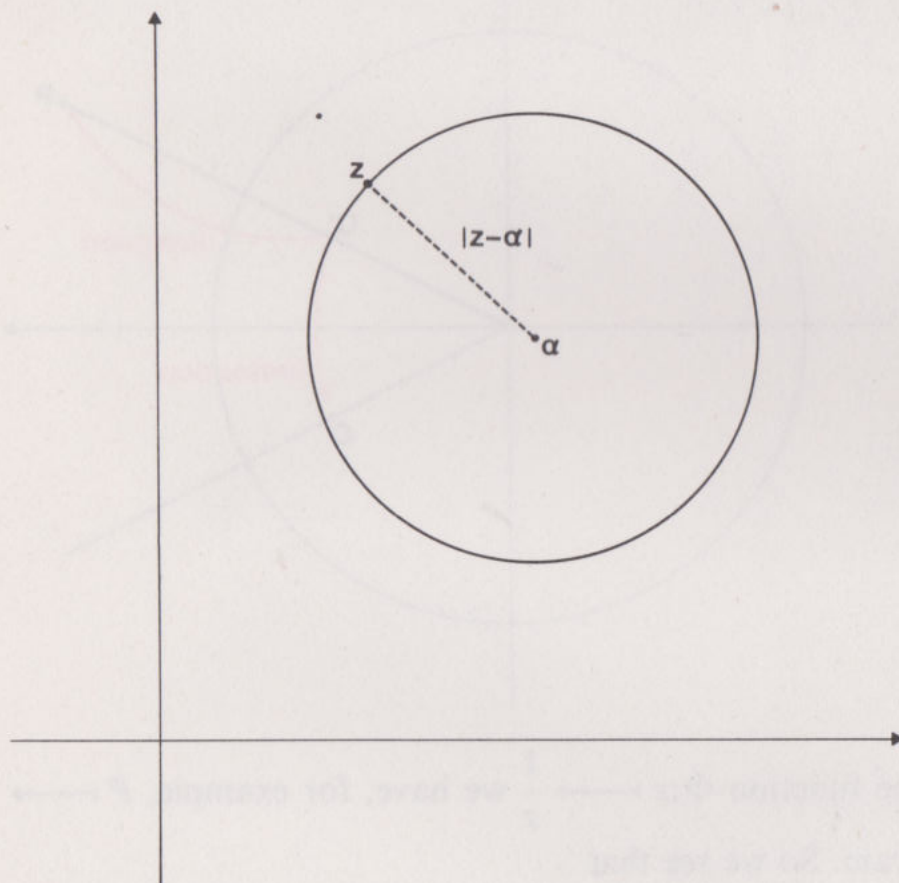
## Exercise 1

Which of the following sets are invariant under the function  $\Phi$ ?

- (i) a straight line through the origin;
- (ii) a straight line parallel to the imaginary axis;
- (iii) a straight line parallel to the real axis;
- (iv) the circle  $\{z: |z| = a\}$ , where  $a \in \mathbb{R}^+$ ;
- (v)  $\{e^{i\theta}, e^{-i\theta}\}$ ;
- (vi)  $\left\{z: \frac{1}{a} \leq |z| \leq a\right\}$ , where  $a \in \mathbb{R}$  and  $a \geq 1$ .

We now investigate algebraically what happens to circles and lines under the function  $\Phi$ . Let's start with circles: a circle with radius  $a$ , and centre at the point representing the complex number  $\alpha$ , has the equation

$$|z - \alpha| = a.$$



We can rewrite this equation as

$$(z - \alpha)(\bar{z} - \bar{\alpha}) = a^2$$

or

$$z\bar{z} - \bar{\alpha}z - \alpha\bar{z} = a^2 - \alpha\bar{\alpha}$$

Under the function  $\Phi: z \mapsto \frac{1}{z}$  this becomes

$$\frac{1}{z\bar{z}} - \frac{\bar{\alpha}}{z} - \frac{\alpha}{\bar{z}} = a^2 - \alpha\bar{\alpha}$$

i.e.

$$1 - \bar{\alpha}\bar{z} - \alpha z = (a^2 - \alpha\bar{\alpha})z\bar{z}.$$

Now if  $a^2 - \alpha\bar{\alpha} \neq 0$ , we can rearrange this as

$$z\bar{z} + \frac{\alpha}{a^2 - \alpha\bar{\alpha}}z + \frac{\bar{\alpha}}{a^2 - \alpha\bar{\alpha}}\bar{z} = \frac{1}{a^2 - \alpha\bar{\alpha}}.$$

If we compare this with the (red) equation of the original circle, we see that it is the same, except that

$$\alpha \text{ has been replaced by } \frac{-\bar{\alpha}}{a^2 - \alpha\bar{\alpha}}.$$

and

$$a^2 - \alpha\bar{\alpha} \text{ has been replaced by } \frac{1}{a^2 - \alpha\bar{\alpha}}.$$

So the image of a circle is a circle, in general.

There is, however, one exceptional case:

$$a^2 - \alpha\bar{\alpha} = 0.$$

What is special about this case? Well,  $\alpha\bar{\alpha} = |\alpha|^2$  and  $|\alpha|$  is the distance of the point representing  $\alpha$  from the origin. So

$$a^2 - \alpha\bar{\alpha} = 0 \text{ implies } a = |\alpha|.$$

i.e. the circle  $|z - \alpha| = a$  passes through the origin. (And if we had wanted to predict this special case we could have done so, since the origin does not belong to the domain of  $\Phi$ . In this case there is, so to speak, a *hole* in our original circle.)

When  $a^2 - \alpha\bar{\alpha} = 0$ , the transformed equation becomes

$$1 - \bar{\alpha}\bar{z} - \alpha z = 0.$$

In the next exercise we ask you to interpret this equation.

### Exercise 2

By writing  $z = x + iy$  and  $\alpha = a + ib$ , determine the set represented by the equation

$$1 - \bar{\alpha}\bar{z} - \alpha z = 0.$$



So we see that the image of a circle under  $\Phi: z \mapsto \frac{1}{z}$  is either a circle or a straight line. We could now go through the same investigation to find the image of a straight line. But with a little cunning there is no need to do this. The function  $\Phi$  is not only one-one, but it is its own inverse. That is, if we apply  $\Phi$  twice, then we end up where we started;

$$\Phi(\Phi(z)) = z.$$

We have seen that

$$\Phi(\text{a circle through the origin}) = \text{a straight line},$$

so that

$$\text{a circle through the origin} = \Phi(\text{a straight line}).$$

We now wish to know *which* straight lines are the images of circles through the origin. This has been covered in Exercise 2; we showed that the circle through the origin, with centre  $\alpha = a + ib$ , is mapped to the line with equation

$$2by - 2ax + 1 = 0.$$

By varying the circle we can vary  $a$  and  $b$  and so obtain “almost any” straight line: the only type of line we cannot get is a line through the origin, because we can never get an equation of the form

$$cy + dx = 0.$$

There is no way of getting rid of the 1! So we know that

$$\Phi(\text{any straight line not through the origin}) = \text{circle through origin}.$$

But we have already shown (Exercise 1, part (i)) that

$$\begin{aligned} \Phi(\text{any straight line through origin}) \\ = \text{some straight line through origin}. \end{aligned}$$

(Notice once again the crucial role of the origin.) This now covers all possible images of straight lines and circles, and we have seen that the image is always a straight line or a circle. This completes our investigation of  $\Phi$ .

## 10.6 Composition of Complex Functions

We can combine complex functions just as we can combine real functions. For example, if

$$f: z \mapsto z + 2 \quad (z \in \mathbb{C})$$



and

$$g:z \longmapsto 3z \quad (z \in C)$$

then

$$f + g:z \longmapsto 4z + 2 \quad (z \in C).$$

Also

$$f \circ g:z \longmapsto 3z + 2 \quad (z \in C)$$

and

$$g \circ f:z \longmapsto 3(z + 2) \quad (z \in C).$$

Often it is very helpful, when considering an apparently complicated function, to break it down into the composite of two or more simple functions.

Let us look at the example

$$F:z \longmapsto \frac{z + 3}{z + 1} \quad (z \in C, z \neq -1).$$

The first step could be to simplify the function to

$$F:z \longmapsto 1 + \frac{2}{z + 1} \quad (z \in C, z \neq -1).$$

Now let us think of the simple processes which would enable us to calculate  $F(z)$  for some arbitrary complex number  $z \neq -1$ .

Using the two functions

$$f:z \longmapsto z + 1 \quad (z \in C),$$

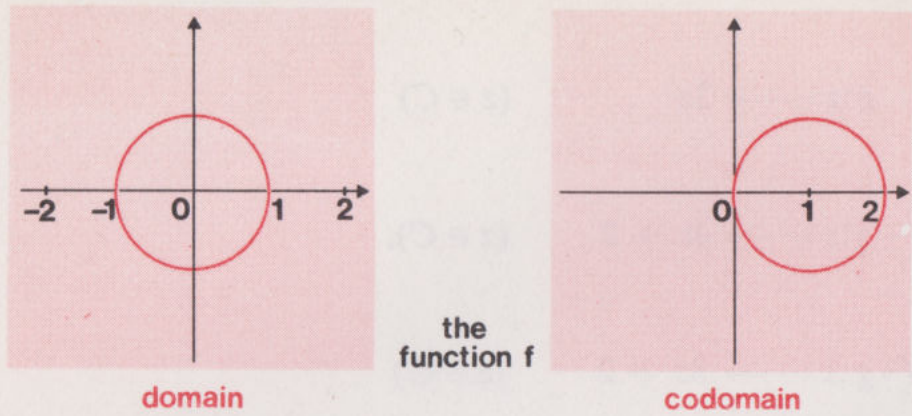
$$g:z \longmapsto \frac{2}{z} \quad (z \in C, z \neq 0).$$

we can write

$$F = f \circ g \circ f.$$

Suppose now that we take a particular set in the domain of  $F$  and attempt to find its image. For example, let us find the image of the circle  $\{z:|z| = 1\}$ . We must of course omit the point  $(-1, 0)$  from the set since this point does not belong to the domain of  $F$ .





First we carry out the translation  $f$ , which simply moves the circle one unit to the right. Now take the new circle which is in the domain of  $g$ . (Notice that this time it is the point  $(0, 0)$  which is omitted from the circle.) The mapping  $g$  can itself be regarded as the composite  $g_2 \circ g_1$  of

$$g_1: z \mapsto \frac{1}{z} \quad \text{and} \quad g_2: z \mapsto 2z.$$

We know what happens under  $z \mapsto \frac{1}{z}$  from our studies in the previous section. The circle passes through the origin, so it is mapped to the straight line with equation

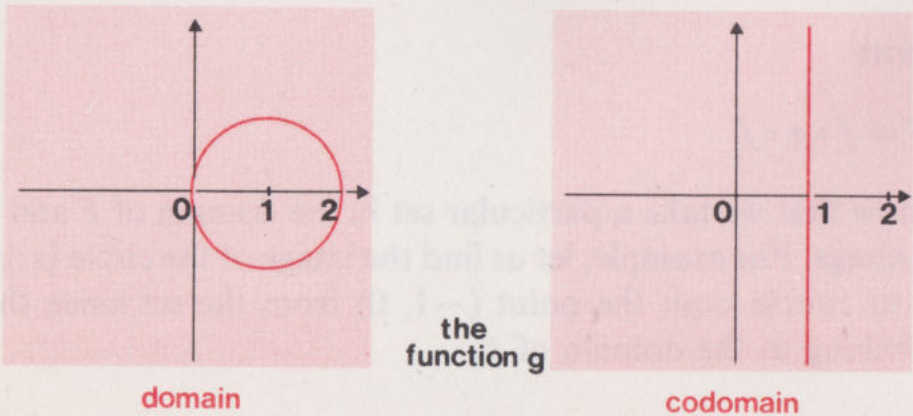
$$2by - 2ax + 1 = 0,$$

where  $(a, b)$  are the Cartesian co-ordinates of the centre of the circle. In this case the centre is  $(1, 0)$ , so the image line has equation

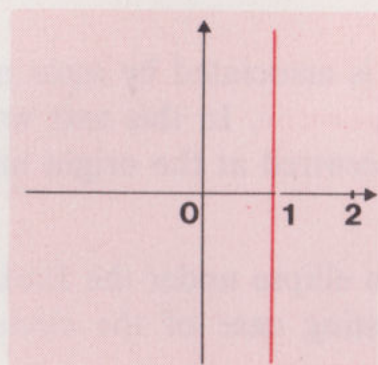
$$-2x + 1 = 0.$$

This is the line parallel to the imaginary axis passing through  $(\frac{1}{2}, 0)$ . This line is then mapped by the scaling  $z \mapsto 2z$ , which doubles the distance of any point from the origin. So the image of the line is the line with equation

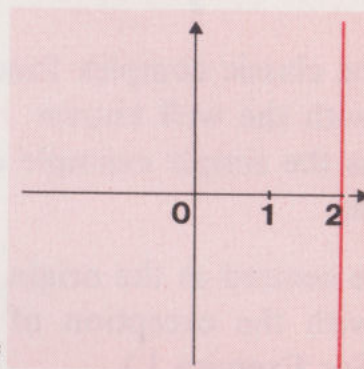
$$x = 1.$$



The next step is to find the image of this straight line under the function  $f$ . Since  $f: z \mapsto 1 + z$ , we simply add 1 to each element in the domain, and therefore the line is moved one unit to the right.

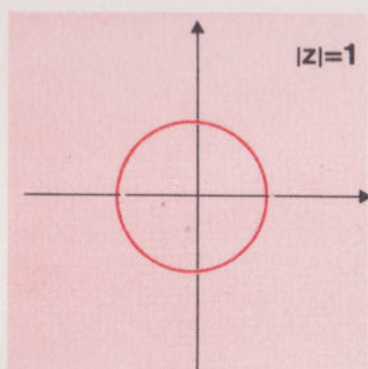


domain

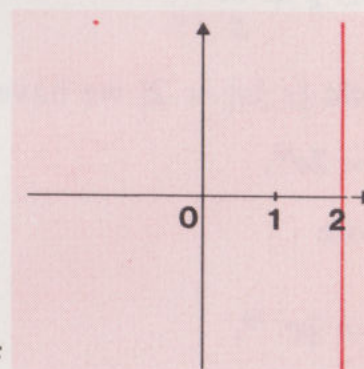
the  
function  $f$ 

codomain

We can now see that the original function  $F = f \circ g \circ f$  maps the circle, with centre the origin and unit radius, to the line parallel to the  $y$ -axis through the point  $(2, 0)$ .



domain

the  
function  $F$ 

codomain

### Exercise 1

Find the image of the circle  $\{z: |z| = 1\}$  under the following functions:

(i)  $F: z \mapsto \frac{3z}{z-1} \quad (z \in \mathbb{C}, z \neq 1)$

(ii)  $G: z \mapsto \left(2 + \frac{3z}{z-1}\right) \quad (z \in \mathbb{C}, z \neq 1)$

(iii)  $H: z \mapsto \left(\frac{3z}{z-1} - \frac{1}{2}\right)^2 \quad (z \in \mathbb{C}, z \neq 1)$



## 10.7 The Joukowski Function

The function

$$z \longmapsto z + \frac{1}{z} \quad (z \in \mathbb{C}, z \neq 0)$$

is one of the classic complex functions and is associated by most mathematicians with the well known **Joukowski aerofoil**. In this text we shall only discuss the simple example of a circle centred at the origin mapped to an ellipse.

Every circle centred at the origin maps to an ellipse under the Joukowski function, with the exception of the interesting case of the circle with radius 1. (See Exercise 1.)

Consider the circle

$$\{z: |z| = 2\}$$

in the domain, and, as before, let  $w = u + iv$  be the image of  $z$ , so that

$$w = z + \frac{1}{z}.$$

On the circle  $\{z: |z| = 2\}$  we have

$$z = 2e^{i\theta},$$

and therefore

$$\frac{1}{z} = \frac{1}{2}e^{-i\theta}.$$

It follows that

$$\begin{aligned} w &= z + \frac{1}{z} \\ &= 2e^{i\theta} + \frac{1}{2}e^{-i\theta} \\ &= 2(\cos \theta + i \sin \theta) + \frac{1}{2}(\cos \theta - i \sin \theta) \\ &= (2 + \tfrac{1}{2}) \cos \theta + i(2 - \tfrac{1}{2}) \sin \theta. \end{aligned}$$

We can now see that

$$u = \tfrac{5}{2} \cos \theta$$

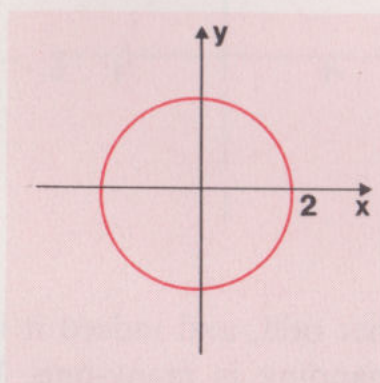
and

$$v = \tfrac{3}{2} \sin \theta.$$

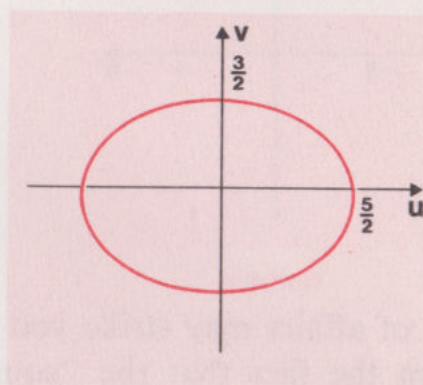
and hence that

$$\frac{4u^2}{25} + \frac{4v^2}{9} = 1,$$

which is the equation of an ellipse with semi-major and semi-minor axes of lengths  $\frac{5}{2}$  and  $\frac{3}{2}$  respectively.



domain



codomain

### Exercise 1

What is the image of the circle

$$\{z:|z| = 1\}$$

under the Joukowski function?

## 10.8 The “Square” Function Again

We have already mentioned some of the properties of the representation of the “square” function in section 10.2, but in this section we wish to examine it again in more detail, in preparation for the final section of this text.

Let us begin by attempting to find the image of the circle  $\{z:|z| = 1\}$  under the “square” function

$$z \longmapsto z^2 \quad (z \in \mathbb{C}).$$

The restriction  $|z| = 1$  on the values of  $z$  in the domain is equivalent to the restriction

$$|z|^2 = 1 \quad \text{or} \quad |z^2| = 1.$$

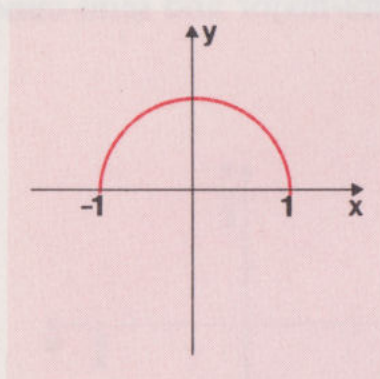
If we let  $w = z^2$  then the corresponding restriction on the values of  $w$  is

$$|w| = 1.$$

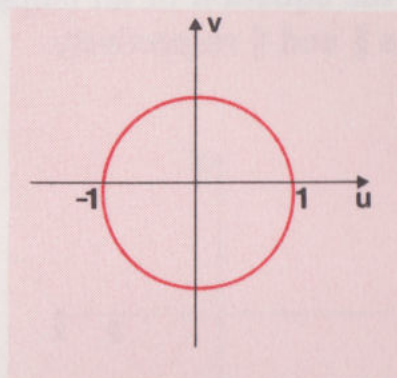
The image set is, therefore, the circle  $\{w:|w| = 1\}$ .



This is not, however, the whole story. We have already seen in section 10.2 that the semi-circle in the upper half of the domain also maps to this circle in the codomain.



domain



codomain

This state of affairs may strike you as rather odd, and indeed it is. It all stems from the fact that the “square” mapping is many-one. We can see more clearly what is happening if we use polar co-ordinates.

We know that the “square” function can be represented in polar co-ordinates by

$$(r, \theta) \longmapsto (r^2, 2\theta) \quad ((r, \theta) \in R_0^+ \times R),$$

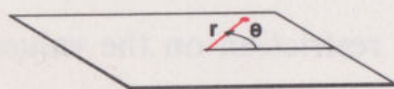
so that the argument of each point in the domain is doubled in the codomain.

As  $z$  moves clockwise round the circle in the domain  $w$  moves twice round the circle in the codomain, also in a clockwise direction.

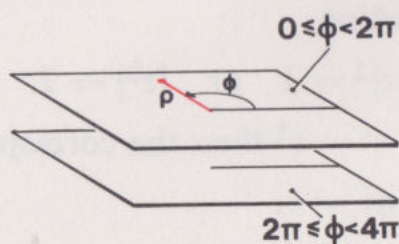
In order to get the full flavour of the function in this case we examine the behaviour of the corresponding polar co-ordinates. If we let  $(\rho, \phi)$  be polar co-ordinates in the codomain, then

$$(\rho, \phi) = (r^2, 2\theta).$$

If we restrict  $\theta$  to lie between 0 and  $2\pi$ , it will imply that  $\phi$  lies between 0 and  $4\pi$ . We can show this behaviour quite clearly if we assign one sheet for the polar co-ordinates of the domain and two sheets for the polar co-ordinates of the codomain.



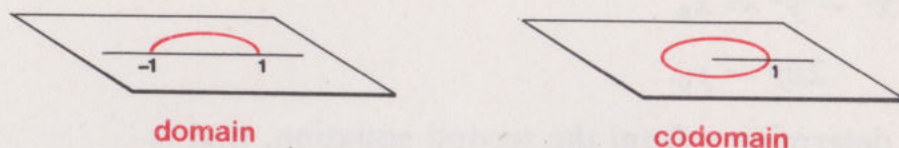
domain



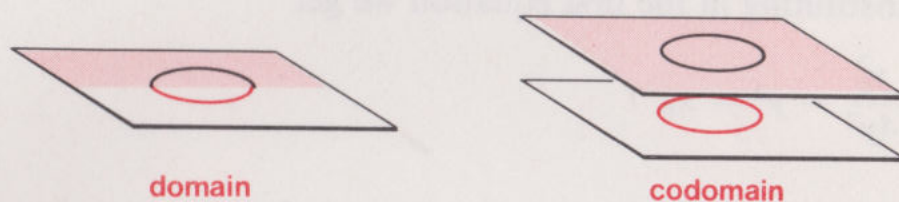
codomain



As a point traces out the semi-circle in the upper half of the domain we obtain a circle on the top sheet of the codomain.



If the point continues round the semi-circle in the lower half of the domain we obtain a circle on the lower sheet of the codomain.



This representation will have advantages when we discuss some reverse mappings in the next section.

### Exercise 1

Show that the image of the line  $x = a$  ( $a > 0$ ) under the “square” function

$$z \mapsto z^2 \quad (z \in \mathbb{C})$$

is a parabola. If a point moves up this line in the direction of the positive  $y$ -axis, how do the image points move?

### Square Roots

In Chapter 9 we were interested in finding the solutions of the equation

$$z^2 = -1,$$

and we discovered that  $i$  and  $-i$  are the “square roots” of  $-1$ . But how do we find the square roots of an arbitrary complex number?

Let us try the most obvious approach. Suppose that we are given a complex number  $z_0 = x_0 + iy_0$  and we wish to find its square roots. We could attempt to solve the equation

$$z^2 = z_0,$$

that is,

$$(x + iy)^2 = x_0 + iy_0,$$

or

$$x^2 - y^2 + 2ixy = x_0 + iy_0.$$



Equating the real and imaginary parts, we obtain the simultaneous equations

$$x^2 - y^2 = x_0$$

$$2xy = y_0.$$

We can determine  $x$  from the second equation, that is

$$x = \frac{y_0}{2y}.$$

Then substituting in the first equation we get

$$\frac{y_0^2}{4y^2} - y^2 = x_0$$

i.e.

$$4y^4 + 4x_0y^2 - y_0^2 = 0.$$

This is a quadratic equation in  $y^2$ , which we can solve easily, and it will in general give rise to *four*\* values of  $y$ , and there will be four corresponding values of  $x$ . The end result is four number pairs as contenders for the square roots of  $z_0$ .

It seems that this technique is not only inelegant but it also produces *four* numbers of which we have to reject *two*. The two extra “solutions” are introduced when we square the expression for  $x$  in terms of  $y$ .

The conclusion surely is that the obvious approach is not very good and we should try something better. This would be even more apparent if we tried to find cube-roots this way.

The answer is to return to our definition of multiplication as an operation of scaling and rotation. Suppose that  $z_0$  has polar co-ordinates  $(r, \theta)$ ; then we wish to find a complex number  $w_0$  with polar co-ordinates  $(\rho, \phi)$ , say, such that

$$w_0^2 = z_0.$$

The polar co-ordinates of  $w_0^2$  are  $(\rho^2, 2\phi)$  and therefore

$$(\rho^2, 2\phi) = (r, \theta).$$

\* Using the formula for the solution of a quadratic equation, we get

$$2y^2 = -x_0 \pm \sqrt{x_0^2 + y_0^2},$$

and whatever  $x_0$ , since  $\sqrt{x_0^2 + y_0^2} \geq \sqrt{x_0^2} = |x_0|$ , the apparent solution

$$2y^2 = -x_0 - \sqrt{x_0^2 + y_0^2} \leq 0$$

is no solution since  $y$  is real and so  $y^2$  cannot be negative.



We might conclude that\*

$$\rho = \sqrt{r} \quad \text{and} \quad \phi = \frac{\theta}{2},$$

and therefore the complex number with polar co-ordinates  $\left(\sqrt{r}, \frac{\theta}{2}\right)$  is a square root of  $z_0$ .

We have obtained only one square root, but after the next exercise we shall discuss a simple way of getting the second.

### Exercise 2

Find polar co-ordinates for the complex number  $1 + i\sqrt{3}$ , and hence find one of its square roots.

The reason why we only get the one square root is that we inferred from the equation

$$w_0^2 = z_0$$

that

$$(\rho^2, 2\phi) = (r, \theta),$$

which is quite unjustified. We know that the polar co-ordinate representation of a complex number is *not unique*, and, before we can write down such an equation, we must at least verify that  $\theta$  is chosen so that the values of  $\phi$  obtained cover all solutions for  $w_0$ . In fact, the non-uniqueness, which has only been a nuisance up till now, can be put to good use in this context.

Suppose  $\theta = \text{Arg } z_0$ ; then one solution of  $w_0^2 = z_0$  can be obtained from

$$(\rho^2, 2\phi) = (r, \theta).$$

But we could just as well choose  $\theta + 2k\pi$  ( $k \in \mathbb{Z}$ ), instead of  $\theta$ . And if, for instance, we choose  $\theta + 2\pi$ , we get

$$(\rho^2, 2\phi) = (r, \theta + 2\pi)$$

whence

$$\phi = \frac{\theta}{2} + \pi.$$

Now we have two solutions

$$\left(\sqrt{r}, \frac{\theta}{2}\right) \quad \text{and} \quad \left(\sqrt{r}, \frac{\theta}{2} + \pi\right)$$

\* When we write  $\rho = \sqrt{r}$  we mean the positive square root, since  $\rho = |w_0|$  is positive.



and these represent *different* complex numbers. Let us just have a look at this in terms of the numerical example in the previous exercise.

In that exercise, it is true that  $1 + i\sqrt{3}$  has polar co-ordinates  $\left(2, \frac{\pi}{3}\right)$  but it also has polar co-ordinates  $\left(2, \frac{7\pi}{3}\right)$ , for instance. In fact, the elements of any one of the pairs

$$\left(2, \frac{\pi}{3} + 2k\pi\right) \quad (k \in \mathbb{Z})$$

are polar co-ordinates of the complex number  $1 + i\sqrt{3}$ .

If we now attempt to find the square roots, we find that their polar co-ordinates are

$$\left(\sqrt{2}, \frac{\pi}{6} + k\pi\right) \quad (k \in \mathbb{Z}).$$

At first sight it seems that we have swung too far in the opposite direction; instead of just one square root we now appear to have an infinite number. Luckily, all is well, for if we find the corresponding complex numbers, we obtain the set of elements of the form

$$\sqrt{2} \cos\left(\frac{\pi}{6} + k\pi\right) + i\sqrt{2} \sin\left(\frac{\pi}{6} + k\pi\right) \quad (k \in \mathbb{Z}),$$

and this set contains only *two* distinct elements, corresponding to  $k = 0$  and  $k = 1$  say:

$$\frac{\sqrt{3} + i}{\sqrt{2}} \quad \text{and} \quad \frac{-\sqrt{3} - i}{\sqrt{2}}.$$

It is not difficult to see why this is so. If, for instance,  $k = 2$ , then the angle becomes  $\frac{\pi}{6} + 2\pi$  and this gives the same complex number as  $k = 0$ . In general, for any  $k$ ,  $k + 2$  and  $k$  gives the same complex number.

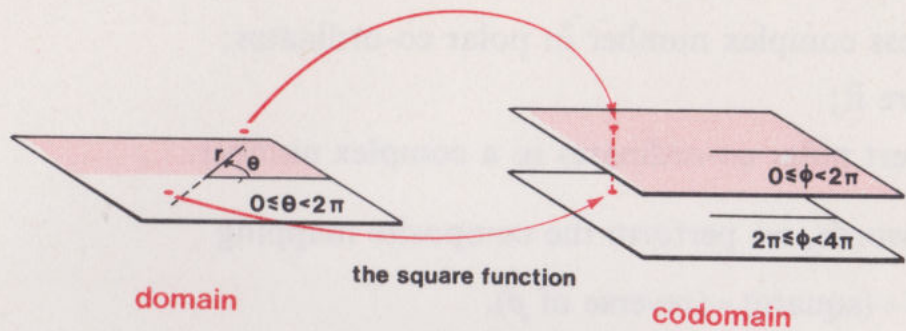
### Exercise 3

Find both the square roots of  $\sqrt{2} + i\sqrt{2}$ .

We shall call the reverse mapping of the “square” function the “**square root**” mapping. This mapping is *not* a function because it is one-many. In order to understand the behaviour of the “square root” mapping, it is helpful to recall what we discovered about the “square” function.



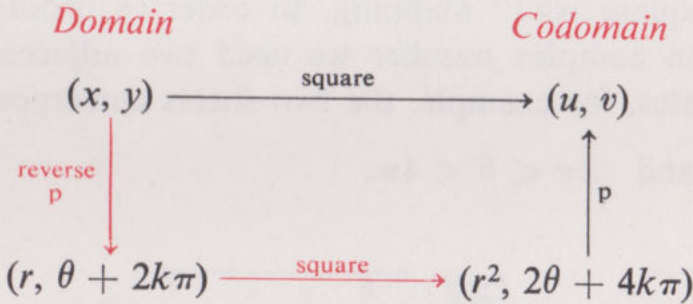
The “square” function maps the set  $\{(r, \theta): 0 \leq \theta < 2\pi\}$  to the set  $\{(r^2, 2\theta): 0 \leq \theta < 2\pi\}$ .



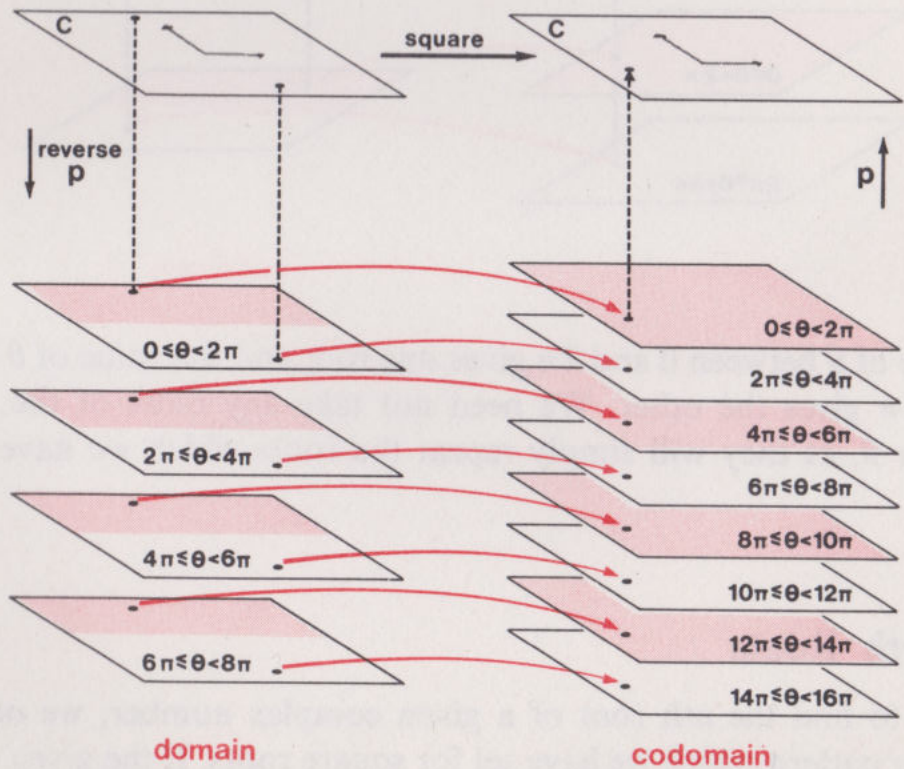
If we now introduce the many-one mapping  $p$  from polar co-ordinates to the corresponding Cartesian co-ordinates:

$$p:(r, \theta) \longmapsto (x, y),$$

we have



If we introduce *all* the values of the argument of a particular complex number, and use a different sheet for each interval  $2k\pi \leq \theta < 2(k + 1)\pi$ , we have diagrammatically:





In order to find the square of a complex number using polar co-ordinates, we could perform the operations:

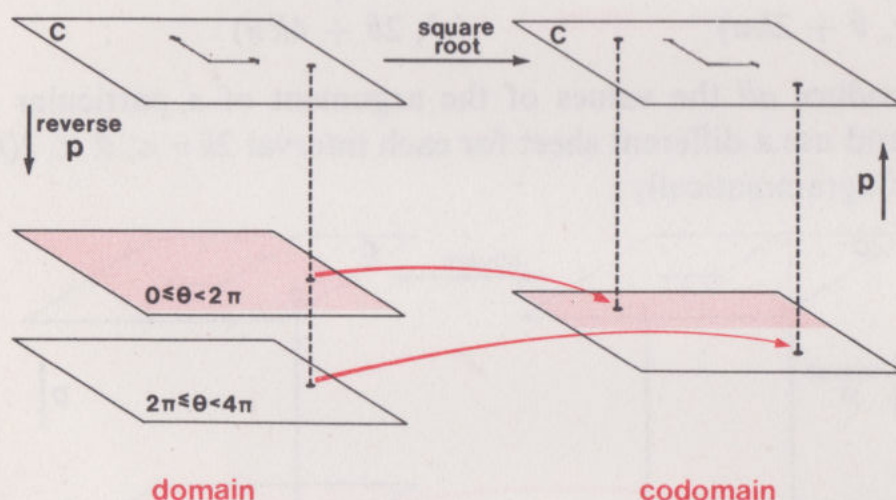
- (i) express complex number in polar co-ordinates;
- (ii) square it;
- (iii) convert polar co-ordinates to a complex number.

In other words, we perform the composite mapping

$$p \circ (\text{square}) \circ (\text{reverse of } p).$$

From the diagram we can see that this process maps two points in the domain to one point in the codomain. In order to produce the image of a particular complex number we need only take values of  $\theta$  lying on the sheet corresponding to  $0 \leq \theta < 2\pi$ . But think now what would happen if we were to proceed round this diagram in the reverse direction as we need to do for the “square root” mapping. In order to produce both square roots of a given complex number we need two adjacent sheets for the polar co-ordinates, for example, the two sheets corresponding to

$$0 \leq \theta < 2\pi \quad \text{and} \quad 2\pi \leq \theta < 4\pi.$$



The value of  $\theta$  between 0 and  $2\pi$  gives one root and the value of  $\theta$  between  $2\pi$  and  $4\pi$  gives the other. We need not take any more of the possible values for  $\theta$ , as they will simply repeat the roots which we have already found.

## 10.9 $n$ th Roots

In order to find the  $n$ th root of a given complex number, we need only follow the pattern which we have set for square roots. If the given complex



number  $z_0$  has  $\text{Arg } z_0 = \theta$  and modulus  $r$ , then  $z_0$  is represented by  $(r, \theta + 2k\pi)$  where  $k \in \mathbb{Z}$ .

Then the  $n$ th roots of  $z_0$  are the solutions  $w_0$  of the equation

$$w_0^n = z_0.$$

If  $w_0$  is represented by  $(\rho, \phi)$  then  $w_0^n$  is represented by  $(\rho^n, n\phi)$ .

From

$$(\rho^n, n\phi) = (r, \theta + 2k\pi),$$

we get

$$\rho = r^{1/n}, \quad \phi = \frac{\theta + 2k\pi}{n} \quad (k \in \mathbb{Z}),$$

whence

$$w_0 = r^{1/n} \left( \cos \left( \frac{\theta + 2k\pi}{n} \right) + i \sin \left( \frac{\theta + 2k\pi}{n} \right) \right).$$

Taking  $n$  successive values of  $k$  will produce the correct number of distinct  $n$ th roots: taking further values of  $k$  will just repeat the roots already found. It is convenient to take  $k = 0, 1, \dots, n-1$ .

### Exercise 1

- (i) Find the values of  $(1 + i)^{1/5}$ .
- (ii) Find the values of  $(1 + i)^{2/5}$ .

### Exercise 2

Solve the equation

$$z^5 = (1 - z)^5.$$

## 10.10 Additional Exercises

### Exercise 1

- (i) Is the complex exponential function an isomorphism or a homomorphism from  $(\mathbb{C}, +)$  to  $(\mathbb{C}_1, \otimes)$ ?
- (ii) What is the image set  $G_1$ ?

(HINT: Some of the results of Exercise 10.4.1 will prove useful.)



*Exercise 2*

Show that all the roots of the equation

$$(z + i)^5 = (z - i)^5$$

are real.

(HINT: There is a quick way!)

*Exercise 3*

Represent the two sets below on an Argand diagram and find the image of each under the function

$$f: z \mapsto 3z - 1 \quad (z \in \mathbb{C})$$

(i)  $\{z: |z + 3| > 3\}$

(ii)  $\{z: \operatorname{Im}(z) = 1\}$

*Exercise 4*

(i) Express the following function as the *composition* of a number of functions

$$g: z \mapsto \frac{3 + 4iz}{1 + 2iz} \quad \left( z \in \mathbb{C}, z \neq \frac{i}{2} \right)$$

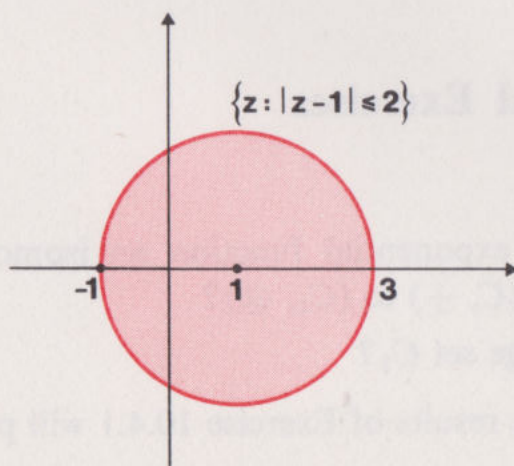
(ii) Hence, find the image of the straight line  $\{z: y = -\frac{1}{4}\}$ .

## 10.11 Answers to Exercises

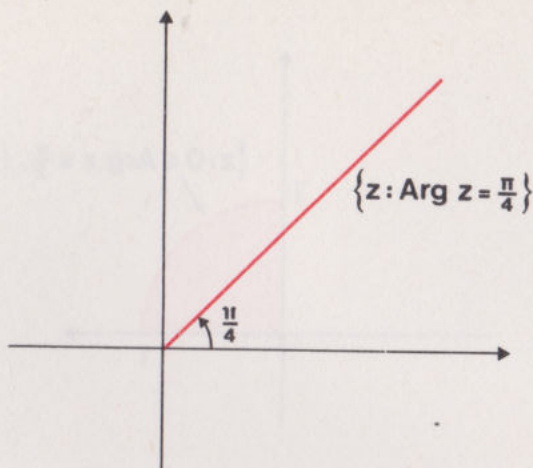
### Section 10.1

*Exercise 1*

(i) “The distance of  $z$  from 1 is less than or equal to 2”.



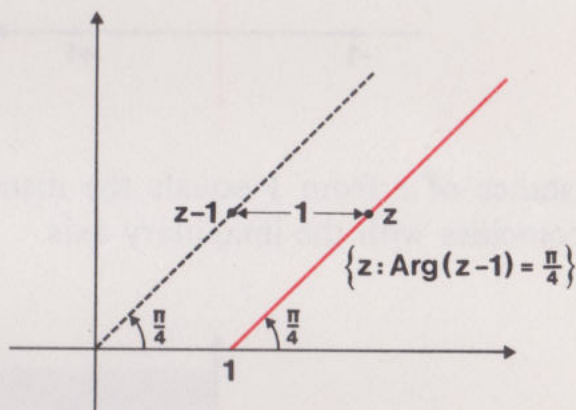
(ii)

(iii) If  $z + 2\bar{z} = 1$  and  $z = x + iy$ , then

$$3x - iy = 1,$$

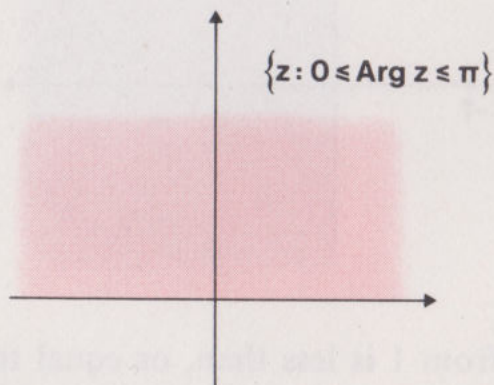
so that  $3x = 1$  and  $y = 0$ . The answer is a diagram showing the single point  $(\frac{1}{3}, 0)$ .

(iv)



The solution is “half line” at an angle  $\frac{\pi}{4}$  to the real axis through the point  $z = 1$ .

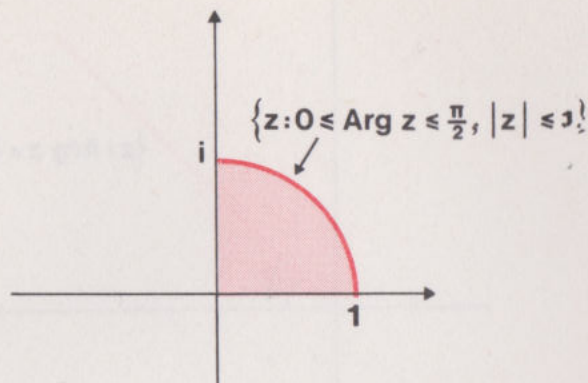
(v)



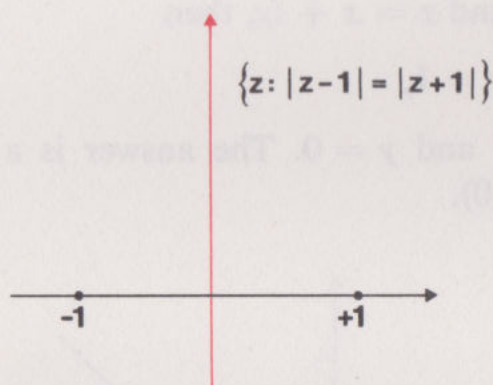
The solution is the upper half-plane including the real axis.



(vi)

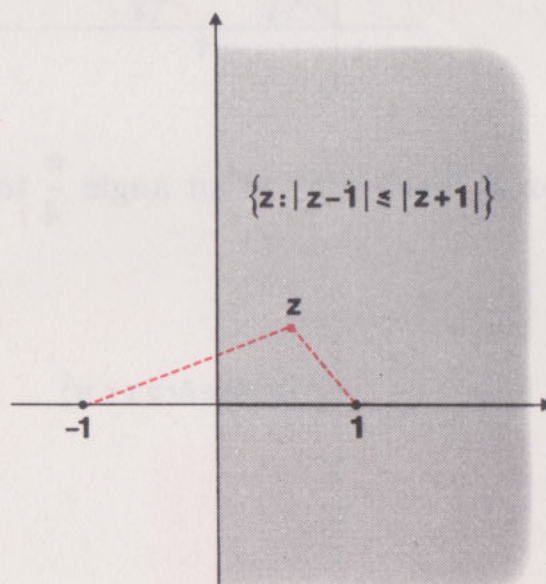


(vii)



“The distance of  $z$  from  $1$  equals the distance of  $z$  from  $-1$ ”, so the set coincides with the imaginary axis.

(viii)



“The distance of  $z$  from  $1$  is less than, or equal to, the distance of  $z$  from  $-1$ ”. The set is the right half-plane including the imaginary axis.

Section 10.2

Exercise 1

We have

$z = iy$  where  $1 \leq y \leq 2$ ,

so

$w = z^2 = -y^2$ ,

i.e.

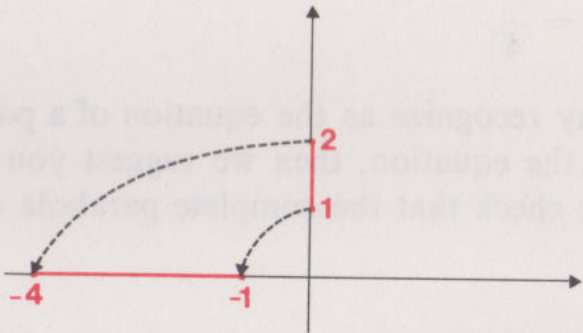
$w = u + iv$

where

$u = -y^2$  and  $v = 0$ .

So the image set is

$\{(u, 0) : -4 \leq u \leq -1\}$ .

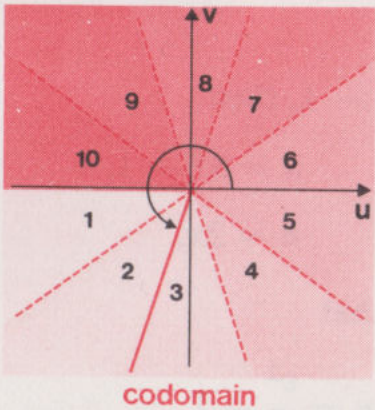
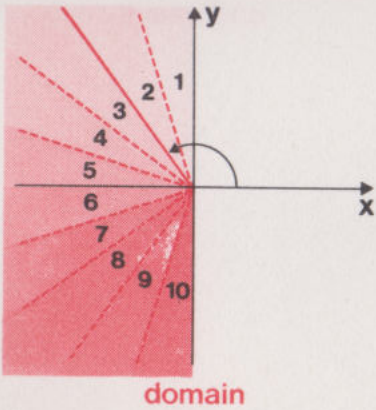


In this diagram the domain and codomain are superimposed.

Section 10.3

Exercise 1

(i)



The image set is the complex plane with the negative real axis (including the origin) removed. We can easily see that this is so if



we notice that the image of every radial line through the origin is also a radial line through the origin but with twice the argument.

(ii) If  $z = x + iy$  and  $w = u + iv$ , then since  $w = z^2$ , we have

$$u = x^2 - y^2,$$

$$v = 2xy.$$

If  $x = 1$ , then

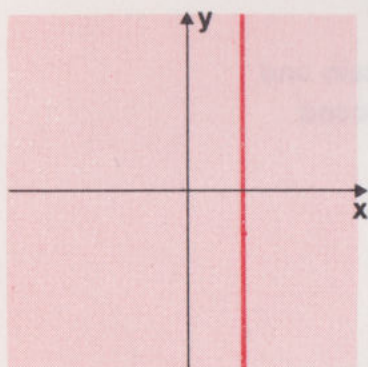
$$u = 1 - y^2$$

$$v = 2y.$$

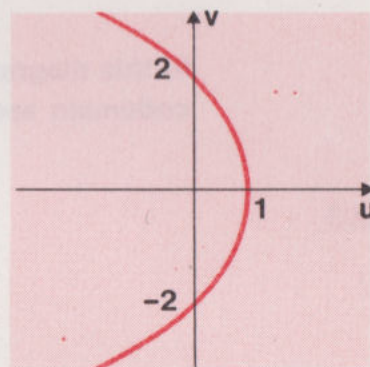
To obtain the image set in the  $w$ -plane, we eliminate  $y$  between these two equations, to obtain

$$1 - u = \frac{v^2}{4},$$

which you may recognize as the equation of a parabola. (If you do not recognize the equation, then we suggest you sketch the graph; you can easily check that the complete parabola corresponds to the line.)



domain



codomain

(iii) If  $y = 1$ , then

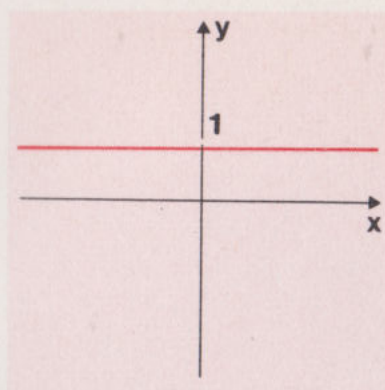
$$u = x^2 - 1,$$

$$v = 2x,$$

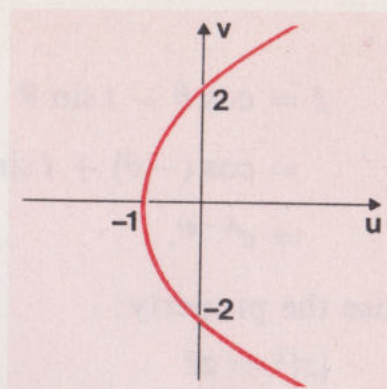
so that

$$u = \frac{v^2}{4} - 1$$

which is again the equation of a parabola.



domain



codomain

**Exercise 2**

Let  $w = 2z + 3$ . The given circle is the set  $\{z: |z| = 1\}$ .

We have

$$z = \frac{w - 3}{2},$$

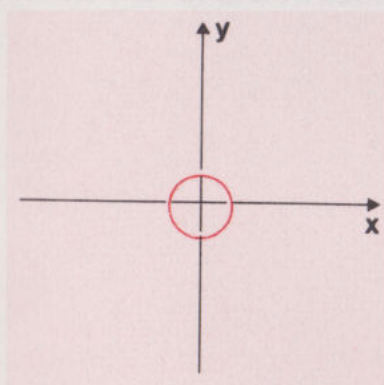
and therefore the image set is determined by the restriction

$$\left| \frac{w - 3}{2} \right| = 1,$$

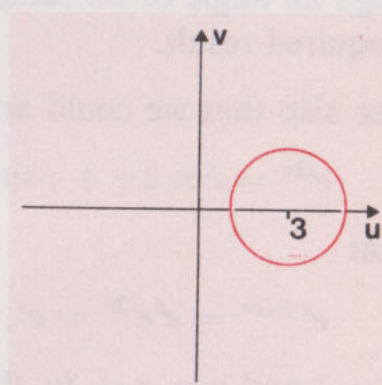
which simplifies to

$$|w - 3| = 2.$$

This means that “the distance of  $w$  from 3 is 2”, so the image set is a circle centred at 3 radius 2.



domain



codomain

**Section 10.4****Exercise 1**

(i) If

$$z = \cos \theta + i \sin \theta,$$



then

$$\begin{aligned}\bar{z} &= \cos \theta - i \sin \theta \\ &= \cos (-\theta) + i \sin (-\theta) \\ &= e^{i(-\theta)}.\end{aligned}$$

(ii) We use the property:

$$\begin{aligned}|z|^2 &= z\bar{z} \\ |e^{i\theta}|^2 &= (\cos \theta + i \sin \theta)(\cos \theta - i \sin \theta), \\ &= \cos^2 \theta + \sin^2 \theta = 1;\end{aligned}$$

hence

$$|e^{i\theta}| = 1,$$

since the modulus of a complex number is non-negative.

$$\begin{aligned}\text{(iii)} \quad |e^z| &= |e^{x+iy}| \\ &= |e^x e^{iy}| \\ &= |e^x| |e^{iy}|.\end{aligned}$$

From (ii),  $|e^{iy}| = 1$ , and since  $|e^x| = e^x$ , we have

$$|e^z| = e^x.$$

$$\text{(iv)} \quad e^{z+i2\pi} = e^z e^{i2\pi}.$$

Multiplying by  $e^{i2\pi}$  corresponds to a rotation about the origin through an angle of  $2\pi$  radians anti-clockwise and this demonstrates the required result.

Notice also that we could argue that

$$e^{i2\pi} = \cos 2\pi + i \sin 2\pi = 1$$

so that

$$e^{z+i2\pi} = e^z e^{i2\pi} = e^z.$$

(v) If  $e^z = 1$  and  $z = x + iy$ , then

$$e^x (\cos y + i \sin y) = 1.$$

Hence

$$e^x \cos y = 1 \quad (\text{a})$$

and

$$e^x \sin y = 0 \quad (\text{b}).$$

From (b) we have  $y = n\pi$ ,  $n \in \mathbb{Z}$ , and substituting in (a) we obtain

$$e^x = \pm 1.$$

But  $x$  is real and there is no real  $x$  such that  $e^x = -1$ , so  $e^x = +1$ .

Hence

$$x = 0.$$

From (a), it follows that

$$\cos y = 1,$$

so  $n$  is even.

Finally,

$$z = i2k\pi, \quad k \in \mathbb{Z},$$

and, just as we would expect, the solution set corresponds to the set of rotations about the origin through multiples of  $2\pi$  anti-clockwise.

### Exercise 2

$$(re^{i\theta})\left(\frac{1}{r}e^{-i\theta}\right) = e^{i\theta}e^{-i\theta}$$

$$= (\cos \theta + i \sin \theta) (\cos (-\theta) + i \sin (-\theta))$$

$$= (\cos \theta + i \sin \theta) (\cos \theta - i \sin \theta)$$

$$= 1.$$

If  $z = re^{i\theta}$ , then it follows that  $\frac{1}{z} = \frac{1}{r}e^{-i\theta}$ .

### Exercise 3

We consider these functions using their geometrical interpretation. They could just as easily be tackled algebraically. For instance, if in (i) we put

$$z = z + z_0,$$

we conclude that *either*  $z_0 = 0$ , in which case every point is invariant, *or* there is no invariant point.

- (i) There are no invariant points unless  $z_0 = 0$ , in which case every point of  $C$  is invariant. Every point  $z$  is translated through a distance  $|z_0|$  in a direction determined by  $\text{Arg}(z_0)$ .



- (ii) If  $\alpha$  is a multiple of  $2\pi$ , then every point of  $C$  is invariant: if not, then the only invariant point is the origin.
- (iii) If  $k = 1$ , then every point of  $C$  is invariant: if not, then the only invariant point is the origin.

#### Exercise 4

Any circle with centre at the origin is invariant under  $z \mapsto e^{i\alpha}z$ , and any straight line through the origin is invariant under  $z \mapsto kz$ .

#### Exercise 5

- (i) If  $z = x + iy$ , then

$$e^{x+iy} = e^{x_1},$$

so that

$$e^x e^{iy} = e^{x_1} (\cos 2n\pi + i \sin 2n\pi).$$

It follows that

$$x = x_1$$

and

$$y = 2n\pi \quad (n \in \mathbb{Z}).$$

- (ii) If  $z = x + iy$ , then

$$e^x e^{iy} = e^{x_1} e^{i\phi},$$

so that

$$x = x_1$$

and

$$y = \phi + 2n\pi \quad (n \in \mathbb{Z}).$$

### Section 10.5

#### Exercise 1

- (i) Since the origin is the centre of inversion, any point on a straight line through the origin maps under a geometric inversion to another point on the same straight line. The reflection of this line in the real axis is a straight line through the origin; the reflection of the line coincides with itself if and only if the line is either the real or the imaginary axis. Only the two axes are invariant.

- (ii) A straight line parallel to the imaginary axis is not invariant, unless that straight line is the imaginary axis itself.
- (iii) A straight line parallel to the real axis is not invariant, unless that straight line is the real axis itself.
- (iv) The image of the set  $\{z: |z| = a\}$  is the set  $\left\{z: |z| = \frac{1}{a}\right\}$ . So the only invariant set is the unit circle.
- (v)  $e^{i\theta}$  and  $e^{-i\theta}$  are images of each other, so the set  $\{e^{i\theta}, e^{-i\theta}\}$  is invariant.
- (vi) This set is an annular region centred at the origin; it is invariant.

### Exercise 2

$$\begin{aligned} 1 - \bar{\alpha}\bar{z} - \alpha z &= 1 - (a - ib)(x - iy) - (a + ib)(x + iy) \\ &= 1 - 2ax + 2by \end{aligned}$$

so that

$$1 - \bar{\alpha}\bar{z} - \alpha z = 0$$

becomes

$$2by - 2ax + 1 = 0,$$

which is the equation of a straight line.

## Section 10.6

### Exercise 1

(i) Let

$$f: z \mapsto z - 1,$$

$$g: z \mapsto \frac{1}{z},$$

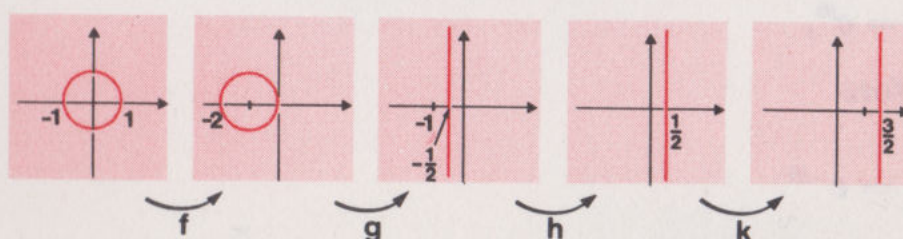
$$h: z \mapsto 1 + z,$$

$$k: z \mapsto 3z,$$

then

$$F = k \circ h \circ g \circ f.$$

Diagrammatically we have:





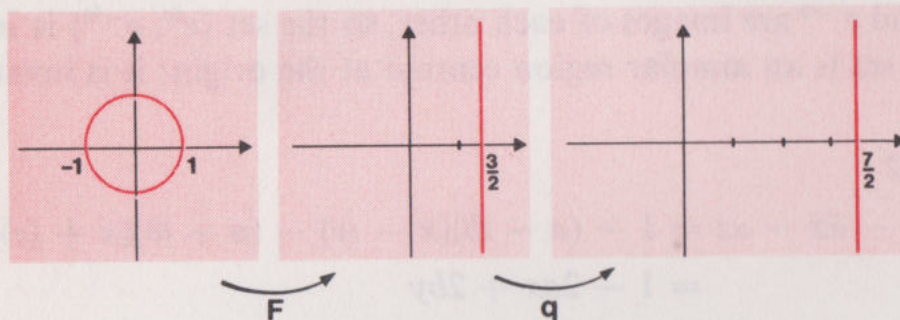
(ii) If

$$q: z \mapsto 2 + z,$$

then

$$G = q \circ F.$$

Diagrammatically we have:



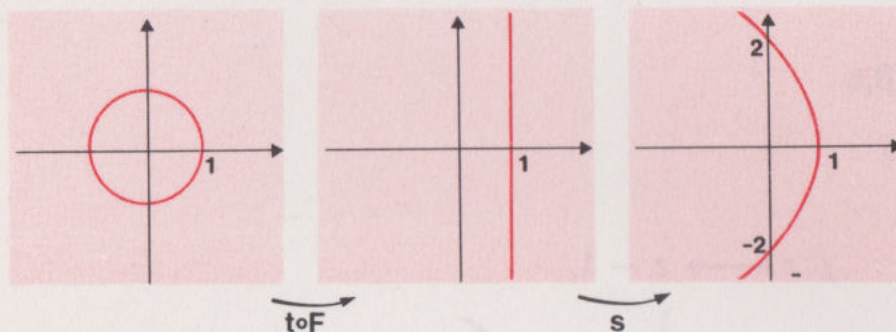
(iii) If

$$s: z \mapsto z^2, \quad t: z \mapsto z - \frac{1}{2},$$

then

$$H = s \circ t \circ F.$$

Diagrammatically we have:



(For the last function see Exercise 10.3.1, part (ii).)

## Section 10.7

### Exercise 1

On the circle  $\{z: |z| = 1\}$  we have

$$z = e^{i\theta},$$

and therefore

$$\frac{1}{z} = e^{-i\theta}.$$

It follows that (in the usual notation)

$$\begin{aligned} w &= e^{i\theta} + e^{-i\theta} \\ &= (\cos \theta + i \sin \theta) + (\cos \theta - i \sin \theta) \\ &= 2 \cos \theta. \end{aligned}$$

Thus

$$u = 2 \cos \theta \quad \text{and} \quad v = 0.$$

As

$$\begin{aligned} \theta &\text{ increases from } 0 \text{ to } \pi, \\ u &\text{ decreases from } 2 \text{ to } -2; \end{aligned}$$

as

$$\begin{aligned} \theta &\text{ increases from } \pi \text{ to } 2\pi, \\ u &\text{ increases from } -2 \text{ to } 2. \end{aligned}$$

We see that the image of  $\{z: |z| = 1\}$  is the real interval  $[-2, 2]$ . ( $w$  traverses this interval twice—once in each direction—as  $z$  travels round the circle once.)

## Section 10.8

### Exercise 1

Let  $w = u + iv$  be the variable in the codomain; then

$$w = z^2$$

and

$$\begin{aligned} u + iv &= (x + iy)^2 \\ &= x^2 - y^2 + 2ixy, \end{aligned}$$

so that

$$u = x^2 - y^2,$$

and

$$v = 2xy.$$

On the straight line in the domain we have  $x = a$ , hence

$$u = a^2 - y^2$$

and

$$v = 2ay.$$

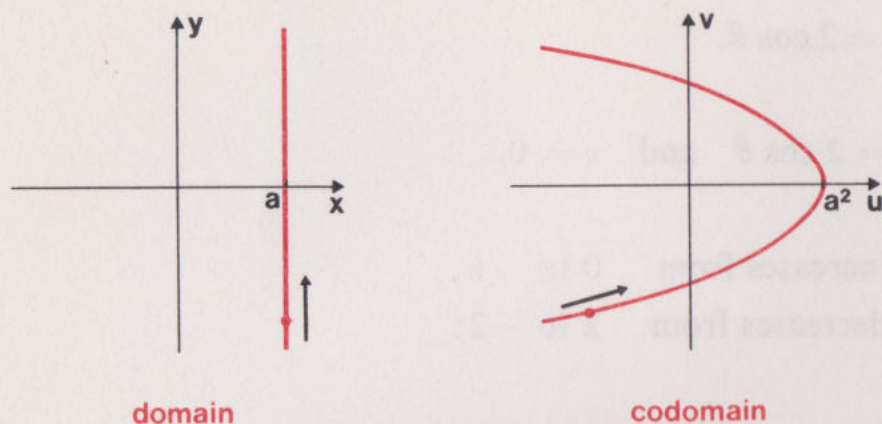
If we eliminate  $y$  between these equations we obtain

$$u = a^2 - \frac{v^2}{4a^2},$$



so that the image set is the parabola

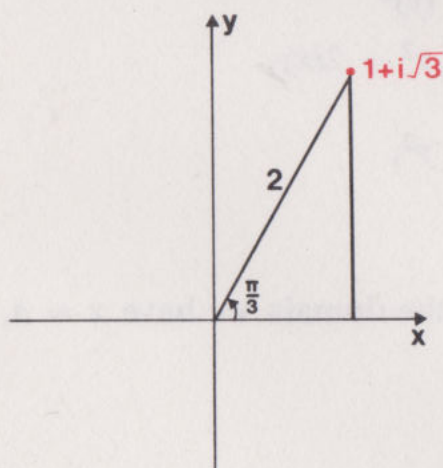
$$\{w: v^2 = 4a^2(a^2 - u)\}.$$



When  $y$  is large and negative, then  $u$  and  $v$  are both large and negative. As the point in the domain moves up the line so the image point moves along the parabola from the lower left, through the vertex and off to the upper left.

The image point crosses the imaginary axis as the point in the domain passes  $(a, -a)$  and  $(a, a)$ , and it crosses the real axis at the vertex of the parabola as the point in the domain crosses the real axis.

### Exercise 2



The complex number  $1 + i\sqrt{3}$  has polar co-ordinates  $\left(2, \frac{\pi}{3}\right)$ ; it follows that the complex number with polar co-ordinates  $\left(\sqrt{2}, \frac{\pi}{6}\right)$  is a square-

root of  $1 + i\sqrt{3}$ . If  $w_0$  is this square root, then

$$\begin{aligned} w_0 &= \sqrt{2} \cos \frac{\pi}{6} + i\sqrt{2} \sin \frac{\pi}{6} \\ &= \sqrt{2} \times \frac{\sqrt{3}}{2} + i\sqrt{2} \times \frac{1}{2} \\ &= \frac{\sqrt{3} + i}{\sqrt{2}}. \end{aligned}$$

### Exercise 3

$\sqrt{2} + i\sqrt{2}$  has polar co-ordinates  $\left(2, \frac{\pi}{4} + 2k\pi\right)$  ( $k \in \mathbb{Z}$ ). The square roots of  $\sqrt{2} + i\sqrt{2}$  therefore have polar co-ordinates

$$\left(\sqrt{2}, \frac{\pi}{8} + k\pi\right) \quad (k \in \mathbb{Z}),$$

and there are just two different square roots:

$$\sqrt{2}e^{i(\pi/8)} \quad \text{and} \quad \sqrt{2}e^{9i\pi/8}.$$

## Section 10.9

### Exercise 1

$$(i) \quad (1 + i) = \sqrt{2} \left( \cos \left( \frac{\pi}{4} + 2k\pi \right) + i \sin \left( \frac{\pi}{4} + 2k\pi \right) \right) \quad (k \in \mathbb{Z}).$$

hence, using the formula in the text, the five values of  $(1 + i)^{1/5}$  are

$$2^{1/10} \left( \cos \left( \frac{\pi}{20} + \frac{2k\pi}{5} \right) + i \sin \left( \frac{\pi}{20} + \frac{2k\pi}{5} \right) \right) \quad (k = 0, 1, 2, 3, 4)$$

(ii) Similarly, the five values of  $(1 + i)^{2/5}$  are

$$2^{1/5} \left( \cos \left( \frac{\pi}{10} + \frac{4k\pi}{5} \right) + i \sin \left( \frac{\pi}{10} + \frac{4k\pi}{5} \right) \right) \quad (k = 0, 1, 2, 3, 4).$$

### Exercise 2

We need to solve the equation

$$\left( \frac{z}{1-z} \right)^5 = 1 \quad (z \neq 1).$$



Let

$$w = \frac{z}{1-z},$$

then the roots of the equation  $w^5 = 1$  are

$$1, e^{2i\pi/5}, e^{4i\pi/5}, e^{6i\pi/5}, e^{8i\pi/5}.$$

If  $w_1$  is a root of  $w^5 = 1$ , then

$$w_1 = \frac{z_1}{1-z_1},$$

where  $z_1$  is a root of the original equation. Solving for  $z_1$  we obtain

$$z_1 = \frac{w_1}{1+w_1}.$$

The roots of the original equation are therefore

$$\frac{1}{2}, \frac{e^{2i(\pi/5)}}{1+e^{2i(\pi/5)}}, \frac{e^{4i(\pi/5)}}{1+e^{4i(\pi/5)}}, \frac{e^{6i(\pi/5)}}{1+e^{6i(\pi/5)}}, \frac{e^{8i(\pi/5)}}{1+e^{8i(\pi/5)}}.$$

## Section 10.10

### Exercise 1

- (i) As a result of Exercise 10.4.1, part (iv), we can conclude that the complex exponential function is many-one. So it is a homomorphism, as opposed to the real exponential function which is an isomorphism.
- (ii) By definition,

$$e^z = e^x (\cos y + i \sin y).$$

Therefore,  $e^z$  can be written as  $(e^x, y)$  in polar co-ordinates. (The modulus of  $e^z$  is  $e^x$ , see Exercise 10.4.1, part (iii), and hence  $y$  is one element of  $\arg(e^z)$ .)

Now  $e^x$  can be any positive number and  $y$  can be any real number, and we shall exhaust all the images by allowing  $y$  to take all values in the interval  $[0, 2\pi[$ . So  $(e^x, y)$  can be the polar co-ordinates of any point in the plane except  $(0, 0)$ . Hence the image set  $C_1$  is  $C$  without the zero element.

### Exercise 2

If

$$(z+i)^5 = (z-i)^5$$

then

$$|z + i|^5 = |z - i|^5,$$

so that

$$|z + i| = |z - i|.$$

All solutions of the original equation are therefore at equal distances from the points corresponding to the complex numbers  $-i$  and  $i$ ; in other words, they lie on the real axis.

### Exercise 3

- (i) The figure shows the set  $\{z: |z + 3| > 3\}$ .

The required region is the *outside* of the circle centre  $-3 + 0i$ , radius 3.

Its image under  $f$  is  $\{w: |w + 10| > 9\}$ , because

$$w = f(z) = 3z - 1, \text{ and hence } z = \frac{w + 1}{3}.$$

Thus, substituting for  $z$  in

$$|z + 3| > 3$$

gives

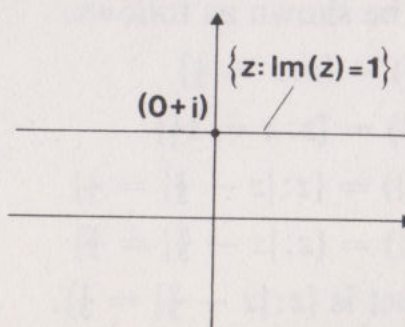
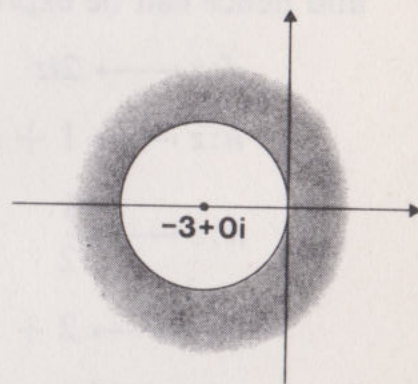
$$\left| \frac{w + 1}{3} + 3 \right| > 3$$

i.e.

$$|w + 10| > 9.$$

The set  $\{w: |w + 10| > 9\}$  is the region *outside* the circle centre  $-10 + 0i$ , radius 9.

- (ii) Since  $\text{Im}(z) = 1$ ,  $z = x + i$ .





If  $f(z) = w$

$$z = \frac{w+1}{3}.$$

Then

$$\operatorname{Im} \left( \frac{w+1}{3} \right) = 1$$

so

$$\operatorname{Im}(w+1) = 3$$

and hence  $\operatorname{Im}(w) = 3$ .

Thus the straight line parallel to the real axis through  $0 + 3i$  is the image under  $f$ .

#### Exercise 4

(i) The function  $g$  can be written

$$z \longmapsto 2 + \frac{1}{1+2iz} \quad \left( z \in C, z \neq \frac{i}{2} \right)$$

and hence can be expressed in terms of

$$f: z \longmapsto 2iz \quad (z \in C)$$

$$h: z \longmapsto 1+z \quad (z \in C)$$

$$k: z \longmapsto \frac{1}{z} \quad (z \in C, z \neq 0)$$

$$l: z \longmapsto 2+z \quad (z \in C)$$

by the composition

$$g = l \circ k \circ h \circ f \quad \left( z \in C, z \neq \frac{i}{2} \right)$$

Notice that other compositions are possible.

(ii) The image of  $\{z: y = -\frac{1}{4}\}$

under  $l \circ k \circ h \circ f$  can be shown as follows.

$$e(\{z: y = -\frac{1}{4}\}) = \{z: x = \frac{1}{2}\}$$

$$f(\{z: x = \frac{1}{2}\}) = \{z: x = 1\frac{1}{2}\}$$

$$g(\{z: x = 1\frac{1}{2}\}) = \{z: |z - \frac{1}{3}| = \frac{1}{3}\}$$

$$h(\{z: |z - \frac{1}{3}| = \frac{1}{3}\}) = \{z: |z - \frac{7}{3}| = \frac{1}{3}\}$$

Therefore the image set is  $\{z: |z - \frac{7}{3}| = \frac{1}{3}\}$ .



# CHAPTER 11 SECOND ORDER DIFFERENTIAL EQUATION

## 11.0 Introduction

It may seem a little unusual to have a chapter on differential equations in an algebra book, but, as we shall see, this chapter employs many of the ideas discussed in the earlier chapters of this volume, as well as ideas from the previous two volumes. Because of the use of so many different concepts introduced elsewhere, this chapter provides a suitable conclusion to the three volumes.

So, in this chapter, we shall draw together several strands of argument that have arisen at various stages of the course, and apply them in the discussion of a familiar physical phenomenon—*resonance*. The word *resonance* refers, in the first instance, to sound: it describes, for example, the acoustic property of bathrooms that makes them flatter our singing voices. However, we shall use the word in a more general sense, to refer not only to sound but also to any kind of vibration, such as the vibrations of an electric motor, or of a car being driven over a bumpy road, or even the electrical vibrations produced in a T.V. or radio set by the electromagnetic waves from the transmitter. We shall describe enough of the physical laws relevant to our problems for the text to be intelligible without prior knowledge of mechanics.

You can easily demonstrate resonance for yourself. Hang any small object, for example, a key, from a string or chain a foot or two in length. Hold the top end of the string or chain in your hand and move it regularly back and forth in a horizontal line a few inches in length. If you do this slowly, the object will follow your hand; if you do it rapidly, the object will hardly move; but if you do it at just the right frequency the object will swing back and forth with a much larger amplitude than that of your hand, and you will find that an almost imperceptible hand movement is enough to keep the object swinging violently.

A more dangerous experiment illustrating the same thing can be done in your bath. By moving your hand (or, if you feel energetic, your whole body) back and forth in the water at just the right frequency you can quickly set up an oscillatory motion big enough to slop water on to the floor.

Another example is that of applying a small intermittent force, at accurately judged times, to a child on a swing, by which you can work him up to quite a big amplitude of motion.



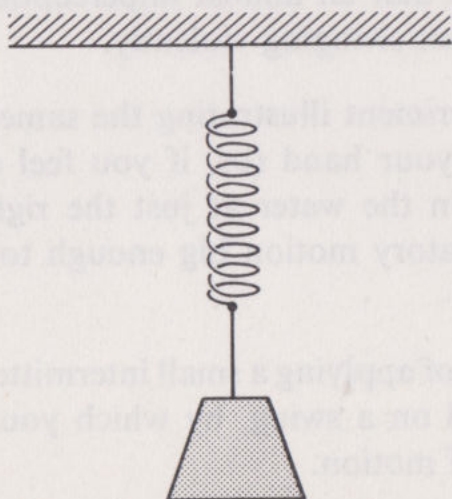
In each of these examples we find that a vibrating system responds very strongly to a force impressed from outside if this impressed force varies with time in just the right way. This strong response is called **resonance**, and we shall study the mathematics of resonance in this chapter.

Before we can use any mathematics at all it is necessary to set up a mathematical representation, or model, of the physical situation we are studying. This mathematical model takes the form of a differential equation, but since this equation is of the second order, the techniques discussed in Volume 2, Chapter 6, cannot be applied directly to yield an exact (or formula) solution. In fact, since there is no deductive method that gives exact solutions for all differential equations, we are really in a “problem-solving” situation: it is necessary to marshal whatever information appears to be relevant and use it to build up a solution to the new type of differential equation. Finally we shall have to examine the solutions we have found and interpret them to obtain an explanation of the physical phenomenon of resonance with which we started.

## 11.1 Setting Up a Model

The first step in setting up a mathematical model for resonance is to simplify the physical situation; then we can more easily see how to describe this situation in mathematical language. The essentials of the physical situation are:

- (i) a system that can oscillate even if it is left undisturbed following an initial disturbance (in the experiments suggested earlier, this is the object on a string, or the water in the bath), and
- (ii) an external to-and-fro disturbance to this system (in the experiments suggested, this disturbance is the motion of your hand).



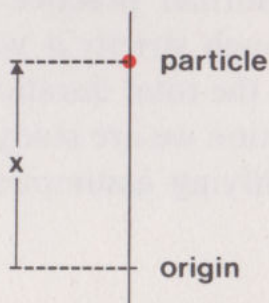


One of the simplest physical systems of this type is the one illustrated in the figure. It consists of a body hanging from a support. If the support is held fixed, the body can oscillate up and down, thus satisfying condition (i) above, and by pushing the body alternately up and down (either by hand or by moving the support up and down) we can apply an oscillatory disturbing force to the system, in accordance with condition (ii) above. The system pictured in the figure is not as artificial as it looks: the support could represent one of the axles of a car, the spring one of the car's springs, and the body the part of the car which is supported by that spring. (This is an over-simplification, of course, because we cannot divide the whole car into four independently moving parts, but it serves to illustrate the way one might set about a preliminary study of the vibrations of a car.)

### Description of the Motion

In setting up a mathematical representation or model of this physical system we are particularly interested in the motion of the body in the vertical direction. The mathematical model consists of two parts: (i) a description of the motion, and (ii) some mathematical equation or equations representing the physical law that determines the motion. So first we shall consider the description of the motion—that is, of how the position of the body depends on time. This is mainly a question of choosing a suitable notation.

To simplify the situation as much as possible, suppose the body is concentrated in a very small region of space, so that it can be represented approximately as a mathematical point. An object so represented is called a **particle**. Since we are studying vertical motion, we assume the particle to move always along a vertical line. We can describe the position of the particle by giving its directed distance from some fixed point (the origin of co-ordinates) on the line. This directed distance is called the **displacement** of the particle; we shall denote it by the variable  $x$ .



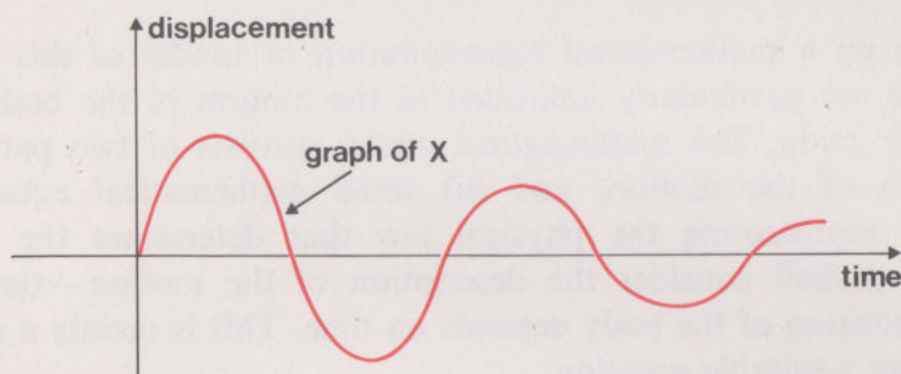
By *directed distance* we mean that the distances on one side of the origin are taken as positive, and those on the other side are taken as negative.



It does not matter which direction we choose to call positive, but it is essential to make a choice and to stick to it consistently. We shall make the (completely arbitrary) choice that the upward direction is positive (i.e. points above the origin have positive values of  $x$ ). Likewise, it does not matter which point we choose as origin. In oscillation problems, however, it is usually convenient to choose the origin to be the *equilibrium position* of the particle—that is, the position at which the particle can remain at rest if undisturbed.

In principle, the motion can be described by giving the position of the particle at every instant of time. Thus the motion is, in principle, described by a list of ordered pairs of the form (*time, position*); in other words, by a real function  $X: \text{time} \mapsto \text{displacement}$ .

The graph of this function is called the *displacement time graph*. The following diagram shows a possible displacement-time graph for an oscillating system.



To measure the time, we once again choose an origin, that is, a particular instant from which time is to be measured. The instant at which the motion begins is often the most convenient origin: with this choice we can define any instant during the motion by giving the time (in, say, seconds) that has elapsed since the motion began. This elapsed time is usually denoted by the variable  $t$ . In this way, the motion is described by a real function  $X: t \mapsto x$  ( $t \in R_0^+$ ), so chosen that the displacement  $x$  at time  $t$  is  $X(t)$  i.e.  $x = X(t)$ .

We have followed the normal practice of taking the domain of this function to be  $R_0^+$ , although strictly it would be more correct to use the interval  $[0, T]$  where  $T$  is the total duration of the conditions determining the particular type of motion we are studying. By choosing  $R_0^+$  as domain we are making the simplifying assumption that these conditions persist for ever.

### Exercise 1

Express the velocity and the acceleration (rate of change of velocity) of the particle at time  $t$  in terms of the derived functions of  $X$ .



*Exercise 2*

What description, analogous to the one in the text, would be suitable for the motion of a rotating piece of machinery, such as the rotor of a lathe?

**Applying the Laws of Mechanics**

Having decided to describe the motion using the function  $X$ , the next step is to find out as much as we can about this function. Mathematics alone tells us nothing about the function; we must appeal to our non-mathematical knowledge about the system under study.

Here this knowledge consists of laws of mechanics that have been discovered experimentally. One of the principal laws in the science of mechanics is **Newton's second law**, which states that **the acceleration of any given particle is proportional to the total force acting on it**. (We shall not enter here into the tricky question of how to define *force*.) In terms of our description of the motion, Newton's second law can be written

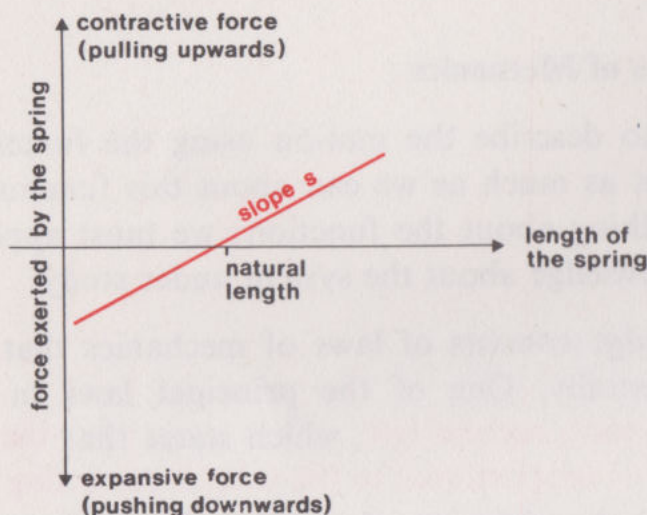
$$X''(t) = kF(t),$$

where  $k$  is the constant of proportionality and  $F(t)$  is the total force acting on the particle at time  $t$ . The heavier the particle, the less it accelerates under a given force, and so the smaller  $k$ : consequently the reciprocal  $1/k$  is larger for heavier particles. This reciprocal is denoted by  $m$  and called the **mass** of the particle. Its numerical value for a given particle depends on the units of measurement used. The standard international unit of mass is the kilogram. We can put the above differential equation in the form  $mX''(t) = F(t)$ .

This equation is the usual form of Newton's second law of motion, but it does not yet tell us anything about the motion; for this we need to know what is meant by the total force  $F(t)$  on the particle—or at least how to calculate it. The “total force” referred to in Newton's law means the sum of the separate forces acting on the particle, and in the present case these include the force of gravity and the force exerted by the spring. For a first look at the behaviour of this system we shall assume that these are the *only* forces acting on the particle. This implies, in particular, that we ignore friction and also (for the time being) deny the possibility of “external” forces such as someone pushing the weight. The motion of a vibrating system which is not affected by any forces (apart from gravity) arising from outside itself is called **free vibration** to distinguish it from motion where an external force (of a particular kind) does act, which is called **forced vibration**.



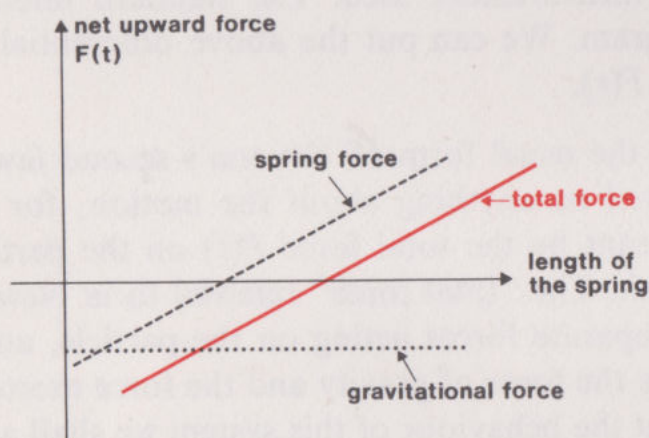
Experiment shows that to a good approximation, for a small system such as that which we are considering, the force of gravity on a particle is a constant and the force of the spring varies linearly with its length (Hooke's Law).



Hooke's Law

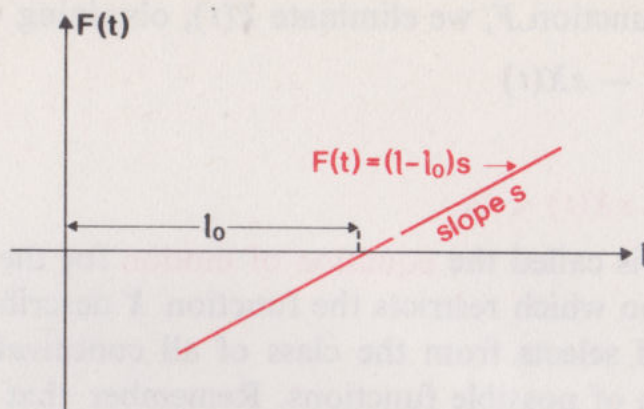
The slope of the line is called the *stiffness* of the spring, and we shall denote it here by  $s$ .

The total force is the sum of the spring force and the downward force due to gravity, which is independent of the length of the spring. Their sum therefore also varies linearly with the length of the spring, and the slope of the new line is again  $s$ .



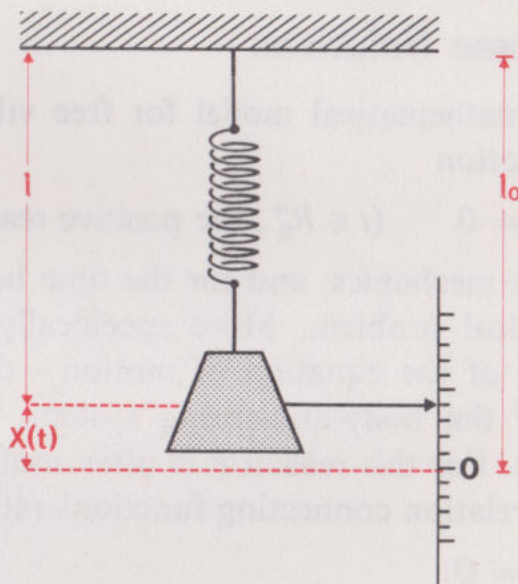
The equilibrium length of the spring is the one at which the two forces on the spring balance out; we denote this length by  $l_0$  and the actual length by  $l$ . We can write the relation indicated in the above diagram between the net upward force  $F(t)$  and the length of the spring, in the form of an equation:

$$F(t) = (l - l_0)s.$$



From the definitions of  $l$ ,  $l_0$  and  $X(t)$ , we have (see the following diagram)

$$l_0 = l + X(t).$$



So  $l - l_0 = -X(t)$  and the formula for  $F(t)$  can be written

$$F(t) = -sX(t).$$

From the science of mechanics we have thus obtained two equations to help us to determine the motion of the particle in the case where the only forces are the spring force and gravity. They are

$$mX''(t) = F(t) \quad (\text{Newton's Law})$$

and

$$F(t) = -sX(t) \quad (\text{Hooke's Law}).$$

These are two simultaneous equations for two unknown functions  $X$  and  $F$ . The usual method for dealing with simultaneous equations is to eliminate one of the unknowns. Here, since we are more interested in the displacement function  $X$ , which describes the motion of the particle,



than in the force function  $F$ , we eliminate  $F(t)$ , obtaining

$$mX''(t) = -sX(t)$$

or equivalently\*,

$$mX''(t) + sX(t) = 0.$$

This last equation is called the **equation of motion** for the system. It is a differential equation which restricts the function  $X$  describing the motion of the system, and selects from the class of all conceivable functions a much smaller class of possible functions. Remember that it is based on many physical assumptions, among which are: the support is fixed; the spring is considered to have negligible weight and to obey Hooke's law exactly; there is no friction and no external force.

## 11.2 Finding Some Solutions

Having set up our mathematical model for free vibrations and arrived at the equation of motion

$$mX''(t) + sX(t) = 0 \quad (t \in R_0^+, \text{ the positive real numbers with zero})$$

we can dispense with mechanics, and for the time being study this equation as a mathematical problem. More specifically, the problem is to find the solution set of the equation of motion—that is, the set of all possible motions of the body-and-spring system. These solutions are *functions*, not images. For this reason it is often useful to write the equation of motion as a relation connecting functions rather than images:

$$mX'' + sX = O.$$

where  $O$  denotes the zero function  $x \mapsto 0$ .

There are various ways of re-writing the equation; for example,

$$X'' + \frac{s}{m}X = O$$

$$D^2X + \frac{s}{m}X = O$$

$$D^2X = -\frac{s}{m}X$$

where  $D^2$  means  $D \circ D$ , i.e.  $D^2: X \mapsto X''$ .

\* Many authors use the notation  $x$  for the displacement  $X(t)$ ,  $\dot{x}$  for the derivative  $X'(t)$  (the velocity) and  $\ddot{x}$  for the second derivative  $X''(t)$  (the acceleration); in this notation the equation of motion is

$$m\ddot{x} + sx = 0.$$



This latter form is interesting: let us restate the problems in words. We are looking for a function  $X$  which when differentiated twice is essentially itself, except for the multiple  $-\frac{s}{m}$ . Or again, in other words, the operator  $D^2$  applied to the function  $X$  produces the same effect as multiplication by the number  $-\frac{s}{m}$ .

Put this way, we might start to look at some of the functions whose derived functions we know, to see if any have this particular form. The process of looking would involve us in selecting a function  $X$ , differentiating it once and then differentiating again. The easiest way to do this is to look at a table of standard derived functions like the one below:

$f$	$f'$	$f''$
$x \longmapsto x^m$	$x \longmapsto mx^{m-1}$	$x \longmapsto m(m-1)x^{m-2}$
$x \longmapsto \exp x$	$x \longmapsto \exp x$	$x \longmapsto \exp x$
$x \longmapsto \ln x$	$x \longmapsto \frac{1}{x}$	$x \longmapsto -\frac{1}{x^2}$
$x \longmapsto \sin x$	$x \longmapsto \cos x$	$x \longmapsto -\sin x$
$x \longmapsto \cos x$	$x \longmapsto -\sin x$	$x \longmapsto -\cos x$

(We have omitted the domains.)

There are many other possibilities but this one will do for now. Three functions in this table are obvious possibilities, namely

$\exp$ ,  $\sin$  and  $\cos$ .

None of them fits exactly, because we have not produced the factor  $-\frac{s}{m}$ , but we can do this by a simple modification:

$f$	$f'$	$f''$
$x \longmapsto \exp(ax)$	$x \longmapsto a \exp(ax)$	$x \longmapsto a^2 \exp(ax)$
$x \longmapsto \sin(ax)$	$x \longmapsto a \cos(ax)$	$x \longmapsto -a^2 \sin(ax)$
$x \longmapsto \cos(ax)$	$x \longmapsto -a \sin(ax)$	$x \longmapsto -a^2 \cos(ax)$



In the case of the exponential function we would have to have  $a^2 = -\frac{s}{m}$ , which means that  $a$  is a complex number: this does not rule out this possibility, but involves some new concepts. So let us look at the other two possibilities. In each case

$$a^2 = \frac{s}{m} \quad \text{so that} \quad a = \pm \omega, \quad \text{where} \quad \omega = \sqrt{\frac{s}{m}},$$

and we have a few possible solutions of our differential equation:

$$\text{or} \quad \left. \begin{array}{l} X:t \longmapsto \cos(at) \\ X:t \longmapsto \sin(at) \end{array} \right\} \quad (t \in R_0^+),$$

$$\text{where } a = \pm \omega \text{ and } \omega = \sqrt{\frac{s}{m}}.$$

(Since  $\cos(-at) = \cos(at)$ , we have three solutions, not four.)

Having arrived at these solutions by various exploratory steps, we should now check that these functions do satisfy the original equation. We leave you to do this, by direct substitution in the differential equation. Another interesting check is to note that we expected to get an oscillatory motion, and that is precisely what the cosine and sine functions represent.

### Exercise 1

Find two solutions of the equation

$$g'' - g = 0,$$

where  $g$  has domain and codomain  $R$ .

### Exercise 2

Find two solutions of the equation

$$g'' - 2g' - 3g = 0,$$

where  $g$  has domain and codomain  $R$ .

### Exercise 3

We have found a few solutions to our equation

$$X'' + \frac{s}{m}X = 0$$



of the particle on the spring. Are there any more? Can you suggest any more? Have you got *all* the solutions?

(We offer no solution to this exercise: the subsequent text is concerned with these questions, but you may like to think about some of these points before reading on.)

## 11.3 Finding the General Solution

The analysis in the preceding section has brought us some way towards the general solution (i.e. the complete solution set) for the equation  $mX'' + sX = 0$  which we have been looking for, but there is still some ground to cover. We have found only a few solutions of the equation of motion, whereas for a full understanding of the free vibrations of the physical system under study we need to be sure that we know *all* the solutions.

There are various ways of going about this problem of generalizing the few solutions we know to obtain the general solution (i.e. the complete solution set). The approach we shall adopt has the advantage of applying to many situations other than the one we are at present considering. What we shall do is to find and make explicit the morphism inherent in our discussion. You will remember that a morphism consists of a function  $h$  from one set to another (or to itself) together with binary operations  $\circ$  and  $\square$ , defined on the domain and image set respectively, satisfying the condition

$$h(x \circ y) = h(x) \square h(y)$$

for all elements  $x$  and  $y$  in the domain of  $h$ . We shall apply the morphism idea in a vector space context to help us with the solution of our present problem.

This is probably not a suggestion which you would consider to be natural, because of the novelty of the ideas involved; yet this is precisely the sort of suggestion which is basic to mathematical thought. We shall try to explain the thought process involved. We are trying to solve an equation; the first thing that is of interest whenever we solve an equation is to be explicitly aware of the set on which the equation is defined. In general, equations can be cast in the form

$$f(x) = b,$$

where  $f$  is some function (occasionally it's a mapping which is not a function, but let's keep it simple),  $b$  is a known element of a codomain of this function and  $x$  (or a set of  $x$ 's) is to be determined in the domain



of  $f$ , which is the set on which the equation is defined. If  $b$  is in the image set of  $f$ , then a solution (or solutions) exist: otherwise they do not exist. Although each type of equation will usually require some special techniques, there are some ideas which apply to many types of equation. We wish to discover and isolate these ideas. To do this we try to determine any special properties of  $f$  and its domain.

Are there any known operations (or relations) on the domain? Is  $f$  a morphism for any of the obvious binary operations on the domain? We have already seen in Chapters 5 and 7 how a vector space structure helps to solve equations.

To put it briefly, the differential equation we are trying to solve is “just another equation”. Its solution will involve some peculiarities because it is a *differential* equation, but it will also involve some general algebraic principles. We have seen some of the peculiarities in the solution functions already found; we are now going to look at some of the general principles.

To return to the present case: we are particularly interested in the set of all functions satisfying the equation  $mX'' + sX = O$ . The characteristic part of our differential equation is the expression  $mX'' + sX$ , and so we are led to study the operator defined by this expression, that is,

$$L:f \longmapsto mf'' + sf \quad (f \in F).$$

(The domain,  $F$ , of  $L$  will be some set of functions. We could stop to specify it more precisely, but, since we are not at present concerned with rigour, it would be a digression to do so.) We can now write our differential equation in the very concise form

$$L(X) = O,$$

but this simplified notation only helps if the relevant properties of the operator  $L$  can also be expressed in a simple form. This is where the morphism comes in.

To satisfy the definition of a morphism, we need binary operations in the domain and image set of the operator  $L$ . The most important binary operations that can be applied to functions were discussed in Chapter 1; they are addition, multiplication and composition. Think for a moment which of these is most likely to produce a morphism when used with  $L$  (i.e. to be compatible with  $L$ ). The rules of differentiation, given in Volume 1, Chapter 3, provide a clue: it was only for addition that we found a morphism, the one described by the rule

$$(f + g)' = f' + g'.$$



This suggests trying addition in both the domain and the image set. We find that

$$\begin{aligned} L(f + g) &= m(f + g)'' + s(f + g) \\ &= (mf'' + sf) + (mg'' + sg) \end{aligned}$$

or, in other words,

$$L(f + g) = L(f) + L(g) \quad (f, g \in F),$$

so that there is indeed a morphism

$$L: (F, +) \longrightarrow (L(F), +)$$

associated with the operator  $L$ .

It would be possible to apply the morphism at once to the solution of our differential equation, but the work will be easier if we follow the algebraic line of thought a little further first. Although we have not specified the domain  $F$  in detail, we can without harm assume that  $F$  is a vector space for the operations of addition of functions and multiplication of functions by real numbers. The morphism property of  $L$  demonstrated above is one of the two morphism properties which characterize a morphism between vector spaces, i.e. a *linear transformation* (see Chapter 5); it is therefore natural to inquire whether  $L$  also satisfies the other morphism property:

$$L(\alpha f) = \alpha L(f) \quad (\alpha \in R \text{ and } f \in F).$$

Using the rules of differentiation, we can verify that this equation does indeed hold, since

$$L(\alpha f) = m(\alpha f)'' + s(\alpha f) = \alpha(mf'' + sf) = \alpha L(f).$$

Thus the operator  $L$  has two properties:

$$L(f + g) = L(f) + L(g)$$

$$L(\alpha f) = \alpha L(f)$$

and therefore satisfies the definition of a vector space morphism; that is,  $L$  is a linear transformation.

Having shown that our differential equation can be written in the form

$$L(X) = O,$$

where  $L$  is a linear transformation, we can use the theory of linear transformations discussed in Chapter 5 to help solve this equation. In fact, we have already studied equations of this form in Chapter 5. We called



the solution set the *kernel* of the linear transformation, and we saw that it had some remarkable properties:

- (i) the kernel is itself a vector space (a subspace of  $F$ );
- (ii) dimension of  $L(F) = (\text{dimension of } F) - (\text{dimension of kernel})$ .

The second property does not help us directly (we have not specified  $F$ ), but the first is very valuable indeed: it tells us that **the solution set of our differential equation is a vector space**. Expressed in plain language, this means that if  $X$  and  $Y$  are any two solutions of the differential equation, then any linear combination of  $X$  and  $Y$  is also a solution of the differential equation.

We are now very close to the solution of our problem, finding the solution set of the differential equation  $mX'' + sX = 0$ . We know a few “solutions” of this equation, namely the functions

$$\left. \begin{array}{l} X:t \longmapsto \cos(at) \\ X:t \longmapsto \sin(at) \end{array} \right\} \quad (t \in R_0^+),$$

where

$$a = \pm \omega \quad \text{and} \quad \omega = \sqrt{\frac{s}{m}}$$

and we also know that the solution set is a vector space. How can these known solutions lead us to a specification of the entire solution set? We saw in Chapter 5 that the most convenient way to specify a vector space is by giving a *basis* of the space, that is, a set of *base vectors* in it such that each element of the vector space can be expressed uniquely as a linear combination of the base vectors. It is natural to guess that perhaps the solutions we already know form a basis for the vector space constituting our solution set—in other words, that every solution of the differential equation is a unique linear combination of the solutions we have already found. There is one snag: a basis is made up of *linearly independent* elements, and of the solutions we have found only two are linearly independent, because

$$\cos(-\omega t) = \cos(\omega t)$$

and

$$\sin(-\omega t) = -\sin(\omega t).$$

So we modify our guess a little and suggest that the solution set of  $mX'' + sX = 0$  is the set of all functions of the form

$$X:t \longmapsto \alpha \cos \omega t + \beta \sin \omega t \quad (t \in R_0^+),$$

where  $\alpha$  and  $\beta$  are arbitrary real numbers, and  $\omega = \sqrt{\frac{s}{m}}$ .



*Exercise 1*

Verify by substitution in the differential equation  $mX'' + sX = 0$  that every function of the above form is a solution of this differential equation.

*Exercise 2*

We have taken it for granted that the two functions we are using for our basis are linearly independent. Can you prove that they are? (Remember that two vectors  $\underline{v}_1$  and  $\underline{v}_2$  are linearly independent if the condition

$$\alpha \underline{v}_1 + \beta \underline{v}_2 = \underline{0}$$

implies that the real numbers  $\alpha$  and  $\beta$  are both 0.)

**A Theorem**

There are still some loose ends to be tied up before we can convincingly claim to have found the general solution of our equation  $mX'' + sX = 0$ . What we have done is to find, by a process of enlightened guesswork, a set of solutions of the equation. Since there is no obvious way of enlarging the set further, it is natural to guess that this set is in fact the set of *all* solutions. Another way of putting it is that we have solved the “problem to find” but we are still left with a “problem to prove”.

Since in these volumes we are more concerned with showing you the main concepts in mathematics than with rigorous proofs, we shall not go into the technique for finding such proofs here; instead we quote a theorem from which the proof we require can be quickly obtained. This theorem applies to a class of operators which includes the specific operator  $L:f \mapsto mf'' + sf$  which we have been considering. The operators in this class are called **linear differential operators**: they are operators of the form

$$L:f \mapsto k_n \times D^n f + k_{n-1} \times D^{n-1} f + \cdots + k_1 \times Df + k_0 \times f$$

( $f \in F_n$ )

where  $n$  is a positive integer called the **order** of the operator,  $F_n$  is a suitable set of functions, and  $k_0, k_1, \dots, k_n$  are real functions with the same domain as the functions in  $F_n$ . (More precisely,  $F_n$  is the set of all functions  $f$ , with codomain  $R$  and domain some subset of  $R$ , for which the  $n$ th derived function is continuous and has the same domain as  $f$ ; the functions  $k_0, \dots, k_n$  are also required to be continuous.) We specify that  $k_n$  is not the zero function (i.e. that its image set contains numbers other than zero); otherwise the term  $k_n \times D^n f$  would be zero, indicating that we had used too large a value of  $n$  in writing down the formula. This require-



ment is introduced to prevent such perversities as writing the operator  $f \mapsto f'$ , whose order is 1, as  $f \mapsto 0 \times f'' + f'$ , which would apparently be of order 2.

As an example, the operator  $f \mapsto mf'' + sf$  is a linear differential operator, with  $n = 2$ ,

$$\left. \begin{array}{l} k_2: t \mapsto m \\ k_1: t \mapsto 0 \\ k_0: t \mapsto s \end{array} \right\} \quad (t \in R_0^+)$$

and  $F_2$  the set of real functions with domain  $R_0^+$  and continuous second derived functions with the same domain. In this case the functions  $k_n, \dots, k_0$  are all constant functions. Whenever all the  $k$ 's are constant functions,  $L$  is called a linear differential operator with **constant coefficients**, and the functions  $k_n, \dots, k_0$  can be treated as numerical multipliers rather than functions. We note (without proof) the following properties:

- (1)  $F_n$  is a vector space for the usual operations:
- (2) any linear differential operator  $L$  of order  $n$  is a vector space morphism (linear transformation) of  $F_n$  to  $L(F_n)$ .

(There is nothing essentially difficult about the proofs of these statements, but they are tedious and not very instructive.)

### Exercise 3

Which of the following operators satisfy the definition of a linear differential operator?

- (i)  $H: f \mapsto D^4f - f \times f$
- (ii)  $H: f \mapsto D^4f - f$
- (iii)  $H: f \mapsto D^4f - (x \mapsto 2)$

In each case the domain is  $F_4$ .

### Exercise 4

Choose one of the operators from Exercise 3 which is not a linear differential operator, and show by means of a counter-example that it is also not a linear transformation.

We are now in a position to state an important theorem.

#### THEOREM

**The dimension of the kernel of a linear differential operator with constant coefficients is equal to the order of the operator.**



Another way of stating the theorem is this: the solution set of an  $n$ th order differential equation of the form  $Lf = O$ , where  $L$  is a linear differential operator with constant coefficients, is a vector space of dimension  $n$ . The theorem also holds for a wide class of cases where  $L$  does not have constant coefficients, but we do not need this generalization here. The proof of the theorem is beyond the scope of this book.

Using this theorem, we can now complete our solution of the differential equation  $mX'' + sX = O$ .

The solution set of this equation is just the kernel of the linear differential operator  $f \mapsto mf'' + sf$ , which is of second order. By the theorem, therefore, the dimension of this kernel is 2: the solution set of  $mX'' + sX = O$  is a vector space of dimension 2. To specify this vector space, all we need is a basis, that is, a set of 2 linearly independent solutions.

We know already that the 2 functions

$$\text{and } \left. \begin{array}{l} t \mapsto \cos \omega t \\ t \mapsto \sin \omega t \end{array} \right\} \quad (t \in R_0^+)$$

are solutions, and that they are linearly independent (Exercise 11.3.2). Consequently they form a basis for the solution space. The solution space itself is the set of all linear combinations of the base vectors, i.e. the set of all functions  $X$  such that

$$X:t \mapsto \alpha \cos \omega t + \beta \sin \omega t \quad (t \in R_0^+),$$

where  $\alpha$  and  $\beta$  are arbitrary real numbers, and  $\omega = \sqrt{\frac{s}{m}}$ .

### Exercise 5

Use the method we have used for the equation  $mX'' + sX = O$  to find all solutions of the differential equation

$$D^2f - f = O,$$

where  $f$  has domain  $R$ .

That is,

- (i) determine the dimensionality of the solution set (a vector space), using the theorem;
- (ii) find enough linearly independent solutions of the equation to form a basis for this vector space;



(iii) write down a formula giving the general solution in terms of these linearly independent solutions.

To end this section we give an example of how some formalistic manipulation can be put to good use.

### Example 1

If we try to solve the equation

$$f'' + 4f' + 5f = 0$$

by the method described in the text, we find that  $t \mapsto \exp(ct)$  is a solution provided that

$$c^2 + 4c + 5 = 0$$

i.e.

$$c = -2 \pm i.$$

This could be considered to be a little embarrassing for at least two reasons:

- (i) we have always assumed that our functions were real functions;
- (ii) we have never discussed differentiating functions which are not real functions, so even if we assume that we have found two solutions, we have no means of verification.

We might, therefore, conclude that if the equation has (real) solutions, they are not of exponential form. That is strictly correct, but there is no need to throw out the baby with the bath water. In Chapter 10 we came across Euler's formula:

$$e^{i\phi} = \cos \phi + i \sin \phi.$$

Using this, we can write

$$e^{(-2+i)t} = e^{-2t}e^{it} = e^{-2t}(\cos t + i \sin t)$$

and

$$e^{(-2-i)t} = e^{-2t}e^{-it} = e^{-2t}(\cos t - i \sin t).$$

Now even if the differentiation of the complex exponential function is in doubt, the algebra is still valid: the vector space spanned by the complex exponential functions  $t \mapsto \exp(-2+i)t$ ,  $t \mapsto \exp(-2-i)t$ , can be spanned by any other two linearly independent vectors in the same space. And we notice that

$$e^{(-2+i)t} + e^{(-2-i)t} = 2e^{-2t} \cos t$$

$$e^{(-2+i)t} - e^{(-2-i)t} = 2ie^{-2t} \sin t.$$



Forgetting about the 2 and  $2i$  (which are just scalar multipliers, the set of scalars in this case being the set of complex numbers), we see that

$$t \longmapsto e^{-2t} \cos t \quad \text{and} \quad t \longmapsto e^{-2t} \sin t$$

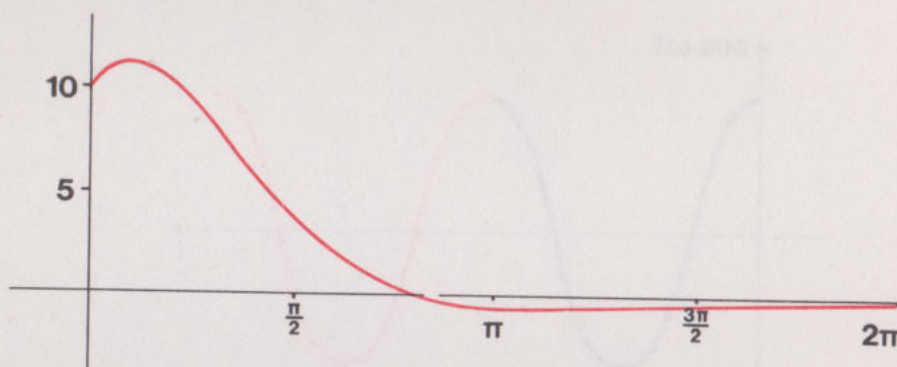
belong to the same two-dimensional vector space. They can be shown to be linearly independent and hence form a basis for that vector space. So we may conjecture that the solution set of the equation is

$$t \longmapsto \alpha e^{-2t} \cos t + \beta e^{-2t} \sin t \quad (t \in \mathbb{R}),$$

where  $\alpha$  and  $\beta$  are arbitrary real numbers. We leave you to verify by substitution that this is correct. The linear independence then means that we have found the set of all real solutions of our equation.

This example illustrates what turns out to be a valid method of solving such equations. It is interesting to note that these differential equations are not the figments of the imagination of a demented mathematician. Such equations occur in electric circuit theory, for instance. The  $e^{-2t}$  has a “damping” effect on the rest of the solution: as  $t$  increases so  $e^{-2t}$  decreases, and for “large”  $t$  the solution becomes effectively

$$t \longmapsto 0 \quad (t \in \mathbb{R}).$$



The above function would arise if, for instance, a system were displaced from its equilibrium position and then left to its own devices without further outside interference. The initial displacement would set up oscillations which would gradually die away, in the same way as the oscillations set up by dropping a stone in water.

## 11.4 Interpretation of the Solution

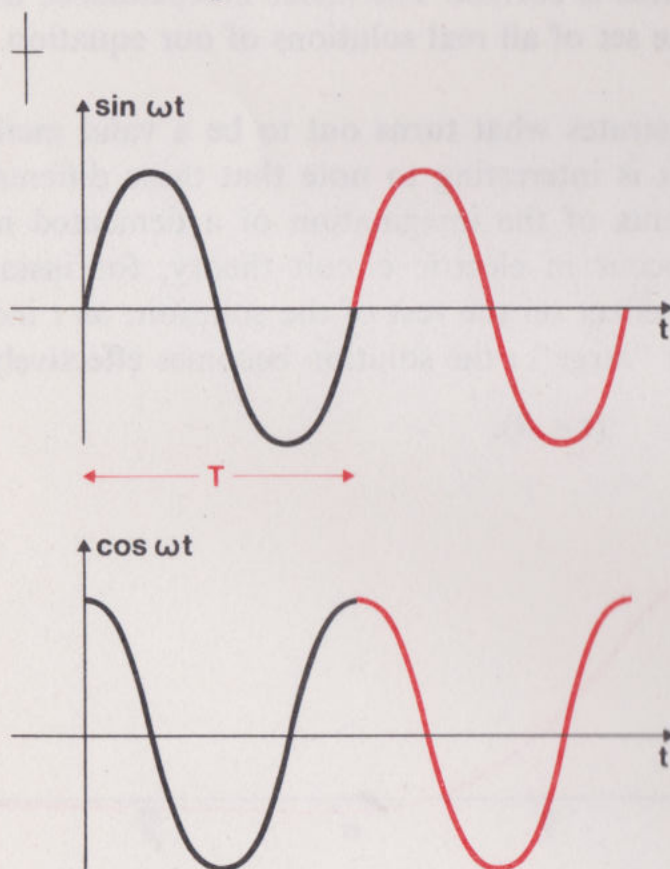
In this section we interpret the general solution,

$$X: t \longmapsto \alpha \cos \omega t + \beta \sin \omega t \quad (t \in \mathbb{R}),$$



which we have found for the equation of motion of our body-spring system, in terms of the possible motions of the system. Every solution of the equation with amplitude less than  $l_0$  corresponds to a possible motion, and so the general solution we have found, with this restriction on  $\alpha$  and  $\beta$ , corresponds to the most general possible motion. We can use it to find the general properties of the motion of the vibrating system.

The most important of these properties is that the motion represented by our solution is *periodic*: it repeats itself exactly again and again. For example, here are the graphs of the two solutions we have been using as base vectors of our vector space of solutions.



Here  $T$  is defined by the equation

$$\omega T = 2\pi$$

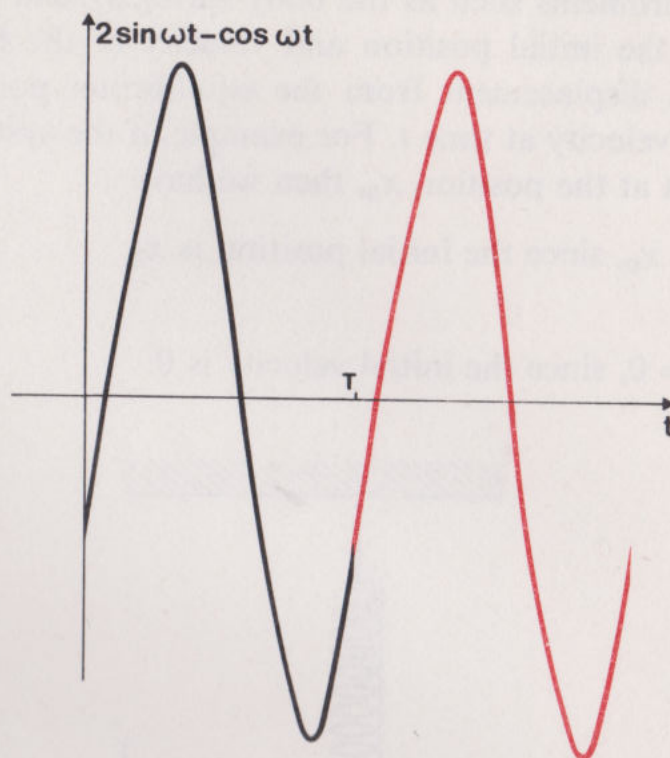
i.e.

$$T = \frac{2\pi}{\omega},$$

so that the first function,  $t \mapsto \sin \omega t$ , maps  $T$  to  $\sin 2\pi$  and the second,  $t \mapsto \cos \omega t$ , maps  $T$  to  $\cos 2\pi$ . In both the graphs the section from 0 to  $T$  is exactly repeated in the sections from  $T$  to  $2T$ ,  $2T$  to  $3T$ , and so on.



Just the same thing happens for linear combinations of the two functions; for example:



The number  $T$  is called the **period** of the motion. In this case  $T$  depends only on the mass of the body and the stiffness of the spring, and not on the particular motion of the vibrating system. The fact that the period remains constant is of the utmost importance for vibrating systems. It ensures, for example, that the pitch of the note emitted by a tuning fork, which is determined by the period of the oscillations of the tines, is not affected by the amplitude (the “size”) of the vibrations—a very desirable prerequisite for tuning forks to fulfil their function of producing notes which serve as a standard of pitch.

Another feature of the general solution which we can relate to the properties of the mechanical system is the presence of two unspecified numbers  $\alpha$  and  $\beta$ , usually called the **arbitrary constants**, in the solution

$$X: t \longmapsto \alpha \cos \omega t + \beta \sin \omega t \quad (t \in \mathbb{R}).$$

For *first-order* differential equations, we found in Volume 2, Chapter 6 that the general solution contained only *one* arbitrary constant. For *second-order* differential equations of the linear type (which is the only type considered here), the general solution has *two* arbitrary constants, and in fact an  $n$ th-order linear differential equation normally has  $n$  arbitrary constants in its general solution.

To pick out, from the family of solutions (the vector space) the particular

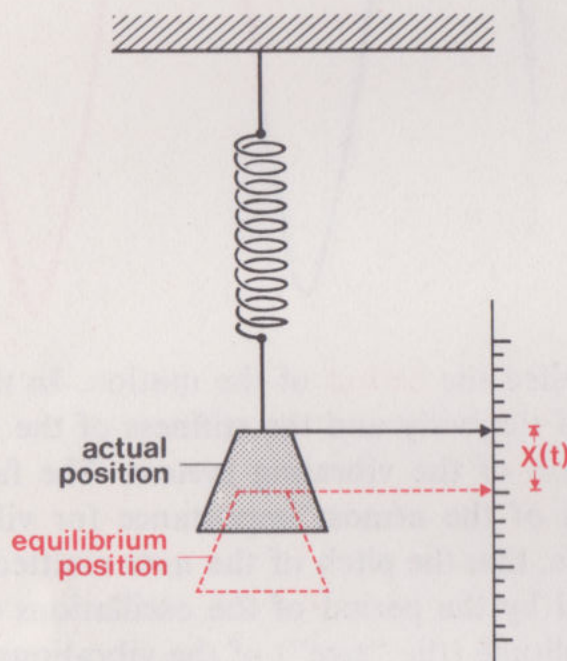


solution describing some particular motion of the mechanical system, we need two pieces of information to determine the two numbers  $\alpha$  and  $\beta$ . In mechanical problems such as the body-spring system this information often concerns the initial position and velocity of the body; remember that  $X(t)$  is its displacement from the equilibrium position at time  $t$ , and  $X'(t)$  is its velocity at time  $t$ . For example, if the system is started at time 0 from rest at the position  $x_0$ , then we have

$$X(0) = x_0, \text{ since the initial position is } x_0,$$

and

$$X'(0) = 0, \text{ since the initial velocity is } 0.$$



These two pieces of information can be used to fix the values of  $\alpha$  and  $\beta$ ; since in this case,  $X(0) = \alpha$  and  $X'(0) = \omega\beta$ , we obtain  $\alpha = x_0$ ,  $\beta = 0$ . So the solution, if the system starts from rest at a position  $x_0$ , is given by

$$X:t \longmapsto x_0 \cos \omega t \quad (t \in \mathbb{R}_0^+).$$

### Exercise 1

If the particle starts at time 0 from the position  $x_0$  with velocity  $v_0$ , find the appropriate solution which describes the displacement.

### Exercise 2

If you wanted to increase the frequency (the number of oscillations per second) of a vibrating system would you



- (a) increase the mass;
- (b) stiffen the spring;
- (c) decrease the mass;
- (d) make the spring less stiff?

## 11.5 A Mathematical Model for Resonance

To deal mathematically with the phenomenon of resonance, we must extend our mathematical model by including the effects of forces other than the force  $-sX(t)$  that arises directly from the displacement of the body. Vibrations in the presence of such additional forces are called **forced vibrations**, to distinguish them from **free vibrations** which take place in the absence of such forces. In studying resonance we are particularly interested in the case where this additional force is periodic: we shall see how the oscillations excited by such a force can build up to a very large amplitude even when the amplitude of the force itself is small. One way in which such an additional force could be transmitted to the system would be for the support of the spring to move in periodic fashion.

For our mathematical treatment of this physical situation, let us take the additional force to have the form  $F_0 \cos pt$ , where  $F_0$  and  $p$  are numbers which we can use to adjust its amplitude and period. The reason for using a cosine or a sine instead of some other periodic function is that the equation of motion is easily solved for a cosine or sine. When this new force is taken into account, Newton's second law gives

$$\begin{aligned} mX''(t) &= \text{total force on body} \\ &= -sX(t) + F_0 \cos pt \quad (t \in R_0^+). \end{aligned}$$

This is usually written with the unknown function  $X$  taken to the left-hand side:

$$mX''(t) + sX(t) = F_0 \cos pt \quad (t \in R_0^+).$$

This equation of motion, with a particular choice of  $p$ , is the basis of our study of resonance. After solving the general case as a mathematical problem, we shall interpret its solution in terms of the behaviour of the physical system it represents.

### Solving the Equation of Motion

We have just dealt with the related problem of solving

$$mX'' + sX = 0,$$



i.e.

$$mX''(t) + sX(t) = 0.$$

We know that the solution set of  $mX'' + sX = 0$  is the vector space spanned by the two functions

$$t \longmapsto \cos \omega t, \quad t \longmapsto \sin \omega t.$$

The equation  $mX'' + sX = 0$  is a special case (with  $F_0 = 0$ ) of the one we are now looking at, and so our problem is to generalize the solution of  $mX'' + sX = 0$  to the case where the right-hand side is a non-zero function.

We could try looking for only a part of the solution set instead of the whole: that is, for a *particular* solution of our equation

$$mX''(t) + sX(t) = F_0 \cos pt.$$

As it happens, this equation does have a simple particular solution. Can you see a way of getting it?

### Exercise 1

Show that, if  $p^2 \neq s/m$ , then there is a number  $k$  such that

$$X: t \longmapsto k \cos pt \quad (t \in R_0^+)$$

is a solution of

$$mX''(t) + sX(t) = F_0 \cos pt.$$

### Exercise 2

Find a particular solution of

$$X''(t) + X(t) = \sin pt, \quad (t \in R_0^+)$$

assuming  $p^2 \neq 1$ .

Coming back to our problem, we now have a particular solution of the differential equation, but this is only a partial solution of our problem since there may be other solutions of the equation.

The equation can be rewritten in various ways. For example, we can use the function notation instead of the image notation. This gives

$$mX'' + sX = t \longmapsto F_0 \cos pt$$

or, more concisely,

$$mX'' + sX = g$$



where  $g: t \longmapsto F_0 \cos pt$ . A still more concise statement would be

$$L(X) = g,$$

where  $L$  is the linear operator that maps  $X$  to  $mX'' + sX$ .

We have come across equations of this latter form before. In section 7.4 we studied systems of linear equations of the form

$$A\underline{x} = \underline{b},$$

where  $\underline{x}$  and  $\underline{b}$  are column vectors and  $A$  is an  $n \times n$  matrix. We found that, if  $A$  is non-singular, then  $A\underline{x} = \underline{b}$  has a unique solution; but that if  $A$  is singular, then  $A\underline{x} = \underline{b}$ , if it has any solution at all, has many solutions. In the latter case, we can find many solutions as follows: if  $\underline{z}$  is any element of the kernel of the mapping represented by  $A$  (that is, a solution of  $A\underline{x} = \underline{0}$ ) and  $\underline{x}_0$  is any solution of  $A\underline{x} = \underline{b}$ , then  $\underline{x}_0 + \underline{z}$  is another solution of  $A\underline{x} = \underline{b}$ ; for we have

$$A(\underline{x}_0 + \underline{z}) = A\underline{x}_0 + A\underline{z} = \underline{b} + \underline{0} = \underline{b}.$$

It is not difficult to prove also that *every* solution of the equation  $A\underline{x} = \underline{b}$  is of the form  $\underline{x}_0 + \underline{z}$  with  $\underline{z}$  in the kernel. (See Exercise 5 below.)

We can now use this analogy. We know that the general solution of the matrix equation

$$A\underline{x} = \underline{b}$$

is

$$\begin{aligned} \underline{x} = & \text{(some particular solution of } A\underline{x} = \underline{b}) \\ & + \text{(general element of the kernel of } A). \end{aligned}$$

We also know that our situation here is effectively the same: the function  $X$  belongs to a set of functions which forms a vector space (corresponding to the vector space to which the  $\underline{x}$ 's belong), and the operator  $L$  is a linear transformation on this vector space (corresponding to the linear transformation represented by  $A$ ). It is therefore to be expected that we can find the general solution of the differential equation  $L(X) = g$  in an analogous way:

$$\begin{aligned} X = & \text{(some particular solution of } L(X) = g) \\ & + \text{(general element of the kernel of } L), \end{aligned}$$

where the kernel of  $L$  is defined as the set of functions that are mapped to  $0$  by  $L$ . A general expression for all the elements of the kernel of a differential operator  $L$  is usually called a **complementary function** of  $L$ , since it "complements" the particular solution to give the general solution.

### Exercise 3

For the case where

$$g: t \longmapsto F_0 \cos pt$$



and

$$L: X \longmapsto mX'' + sX,$$

we have just found a particular solution of  $L(X) = g$ , and we found the general solution of  $L(X) = 0$  earlier in this text. Use this information to write down a formula for the general solution of the differential equation

$$mX'' + sX = t \longmapsto F_0 \cos pt,$$

where

$$p^2 \neq \frac{s}{m}.$$

Having discovered a method that appears to solve the problem, you may be tempted to look no further and to get on as quickly as possible with something else. It is usually a mistake to yield to this temptation: you may fail to make the most of the effort you have spent in finding a method. The supposed solution may be incomplete in some way, or even wrong; further, there may be some useful lesson to be learned from it. For this reason, it is always a good idea to **check the result**. That is, we look back at the original problem and make sure that we really have the solution—neither more nor less.

In the present case, our problem is to find the general solution (i.e. the solution set) of the differential equation:

$$mX''(t) + sX(t) = F_0 \cos pt \quad (t \in R_0^+),$$

and we have arrived at the supposed solution:

$$X(t) = \left( \frac{F_0}{s - mp^2} \right) \cos pt + \alpha \cos \omega t + \beta \sin \omega t,$$

where  $\alpha$  and  $\beta$  are arbitrary real numbers.

The following two exercises, taken together, ask you to check that this really is the general solution. In both exercises you should assume that  $p^2 \neq s/m$ .

#### Exercise 4

Check that every function of the form given above is a solution of the differential equation, whatever values are taken for  $\alpha$  and  $\beta$ .

#### Exercise 5

Check that every real solution of the differential equation has the form given above.



## 11.6 Interpretation of the Solution

It remains to interpret the general solution we have found for the equation

$$mX''(t) + sX(t) = F_0 \cos pt \quad \left(p^2 \neq \frac{s}{m}\right).$$

Our solution has the form

$$X:t \longmapsto \left(\frac{F_0}{s - mp^2}\right) \cos pt + \alpha \cos \omega t + \beta \sin \omega t \quad (t \in R_0^+).$$

As we have seen, the right-hand side is the sum of

(i) a particular solution:

$$t \longmapsto \left(\frac{F_0}{s - mp^2}\right) \cos pt \quad (t \in R_0^+),$$

under which the image is proportional to the impressed force  $F_0 \cos pt$  and contains no arbitrary constants (other than those which describe the impressed force itself);

and

(ii) a complementary function:

$$t \longmapsto \alpha \cos \omega t + \beta \sin \omega t \quad (t \in R_0^+),$$

which does not depend on the nature of the impressed force, and contains two arbitrary constants just as in the case of free vibrations.

Because of the two arbitrary constants  $\alpha$  and  $\beta$ , the expression for  $X(t)$  gives a whole family of functions describing possible motions of the system, one for each pair of numbers  $(\alpha, \beta)$ . To choose the particular member of this family corresponding to a particular motion of the system, we need two pieces of information; these are usually the position and velocity of the vibrating body when the motion starts.

### Example 1

Suppose the particle starts from rest at the origin; that is,

$$X(0) = 0 \quad (\text{starts at origin}),$$

$$X'(0) = 0 \quad (\text{starts at rest}).$$

The general solution gives

$$X(0) = \left(\frac{F_0}{s - mp^2}\right) \cos 0 + \alpha \cos 0 + \beta \sin 0 = \left(\frac{F_0}{s - mp^2}\right) + \alpha$$

$$X'(0) = \left(\frac{-pF_0}{s - mp^2}\right) \sin 0 - \alpha\omega \sin 0 + \beta\omega \cos 0 = \beta\omega,$$



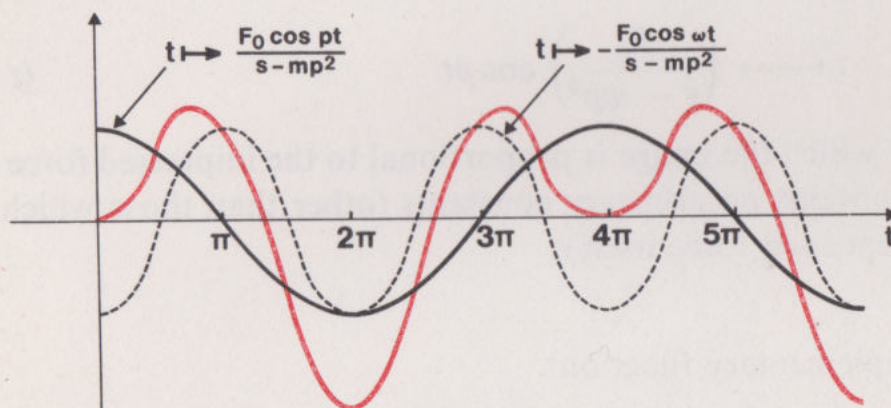
and so using the conditions  $X(0) = X'(0) = 0$ , we obtain

$$\alpha = \frac{-F_0}{s - mp^2}, \quad \beta = 0.$$

The particular solution of the equation appropriate to the given information is therefore

$$X:t \longmapsto \left( \frac{F_0}{s - mp^2} \right) \cos pt - \left( \frac{F_0}{s - mp^2} \right) \cos \omega t \quad (t \in R_0^+).$$

The following graph illustrates the case for which  $F_0 = s = m = 1, p = \frac{1}{2}$ .



### Exercise 1

If the particle starts from a position  $x_0$  with velocity  $v_0$ , find the formula for  $X(t)$  which describes the subsequent motion.

The graph in Example 1 shows that the motion in forced vibrations is likely to be more complicated than in free vibrations. This is only to be expected, because the differential equation for forced vibrations is more complicated. The graph no longer has the character of a sinusoidal curve of period  $2\pi/\omega$  as it does for free vibrations; instead, in general, it consists of a superposition of this sinusoidal wave and another, whose period is determined by the function giving the impressed force. The superposition of these two motions with different periods leads to quite complicated motions (which need not themselves be periodic).

### Exercise 2

(i) What is the period of the function

$$t \longmapsto \cos pt \quad (t \in R)?$$

(ii) What is the period of the function

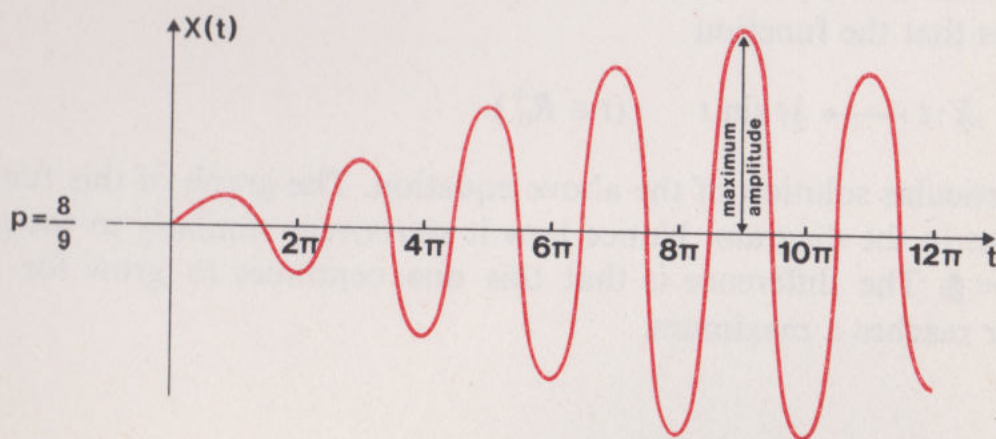
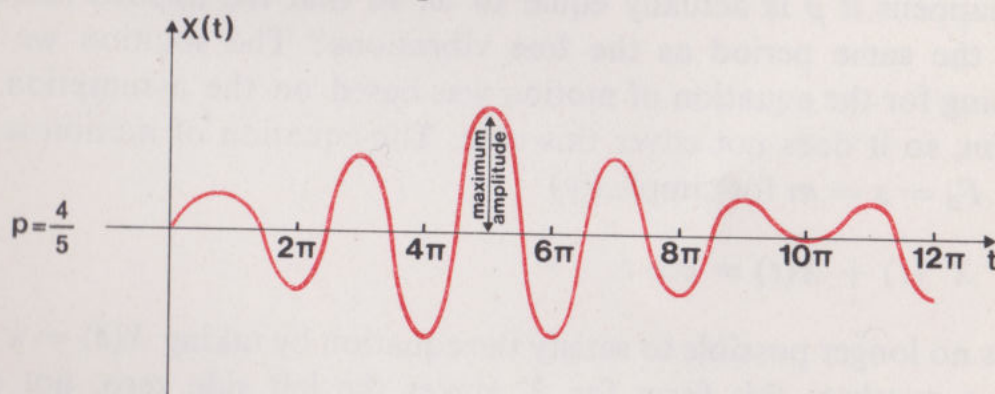
$$X:t \longmapsto \frac{4}{3} \sin \frac{1}{2}t - \frac{2}{3} \sin t \quad (t \in R_0^+)?$$



As indicated at the beginning of this chapter, an important application of the mathematics of vibrations is to the phenomenon of resonance. By *resonance* we mean the tendency of the oscillations of a vibrating system to reach very large amplitudes when the force on it oscillates with a period equal, or very close, to the period of its free vibrations. Our model gives a quantitative explanation of this phenomenon. Suppose, for definiteness, that the impressed force is  $F_0 \cos pt$  and that the system starts from rest at the origin at time 0. Then, the result of Example 1 shows that the subsequent motion is given by

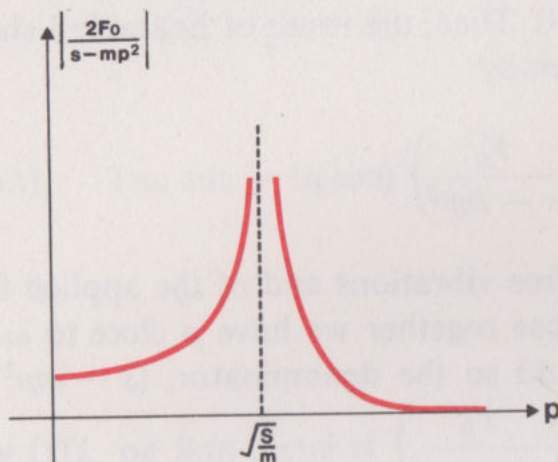
$$X: t \longmapsto \left( \frac{F_0}{s - mp^2} \right) (\cos pt - \cos \omega t) \quad (t \in R_0^+).$$

The periods of the free vibrations and of the applied force are  $2\pi/\omega$  and  $2\pi/p$ . If they are close together we have  $p$  close to  $\omega$  which implies that  $mp^2$  is close to  $s$ , and so the denominator,  $(s - mp^2)$ , is small. Consequently, the factor  $\left( \frac{F_0}{s - mp^2} \right)$  is large, and so  $X(t)$  will be large except at those times (such as the initial time) when the two cosines cancel out, or nearly cancel out. The following diagrams show graphs of  $X$ , with  $F_0 = m = s = 1$ , for two different values of  $p$ . The closer  $p$  is to  $\omega = 1$ , the larger is the maximum amplitude.





To get a simple estimate of the maximum amplitude, let us assume that this is achieved at a value of  $t$  such that one of the two terms  $\cos \omega t$  and  $\cos pt$  is close to  $+1$  and the other is close to  $-1$ . Then  $|\cos pt - \cos \omega t| \simeq 2$  and so the amplitude is about  $\left| \frac{2F_0}{s - mp^2} \right|$ . This quantity is plotted against  $p$  in the next diagram.



We see that the estimated maximum amplitude can indeed be very large when  $p$  is close to  $\omega$ .

What happens if  $p$  is actually equal to  $\omega$ , so that the applied force has exactly the same period as the free vibrations? The solution we have been using for the equation of motion was based on the assumption that  $p^2 \neq s/m$ , so it does not cover this case. The equation of motion is now (taking  $F_0 = s = m$  for simplicity)

$$X''(t) + X(t) = \cos t$$

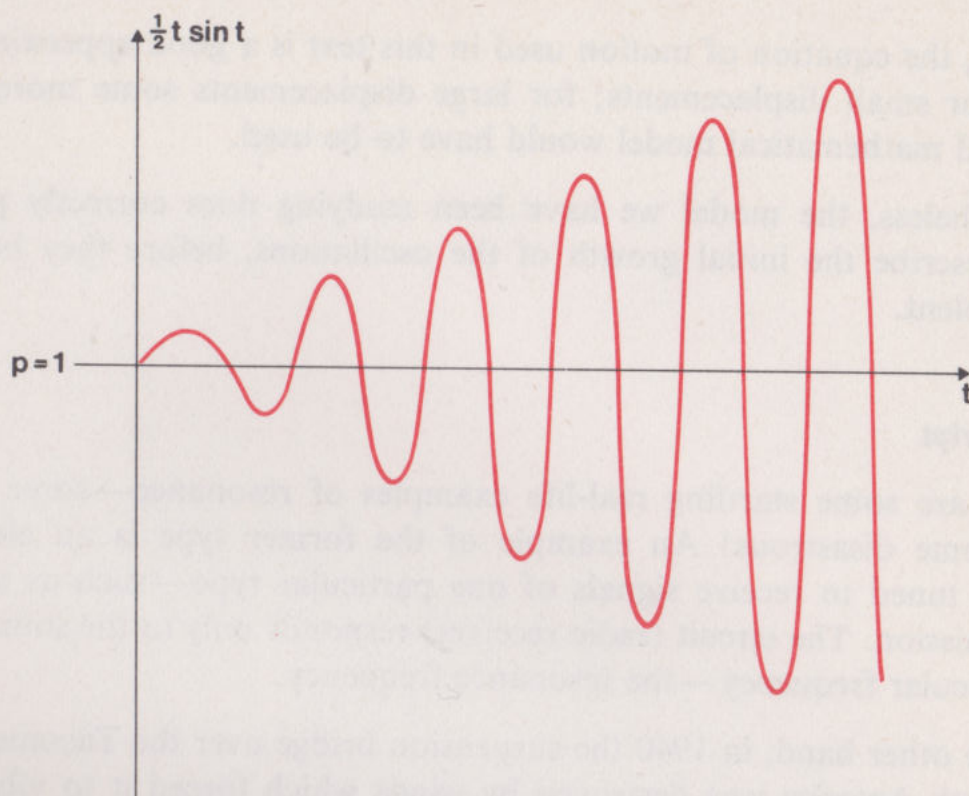
and it is no longer possible to satisfy the equation by taking  $X(t) = k \cos t$  with  $k$  a number: this form for  $X$  makes the left side zero, not  $\cos t$ , whatever  $k$  we use.

There are various methods for finding particular solutions in these exceptional cases. All we are interested in here is the result, however, which is that the function

$$X: t \longmapsto \frac{1}{2}t \sin t \quad (t \in \mathbb{R}_0^+)$$

is a particular solution of the above equation. The graph of this function is shown in the diagram. Notice how it starts very similarly to the graph for  $p = \frac{8}{9}$ . The difference is that this one continues to grow for ever: it never reaches a maximum.



**Exercise 3**

Verify that

$$X: t \longmapsto \frac{1}{2}t \sin t$$

is a particular solution of

$$X''(t) + X(t) = \cos t.$$

**Exercise 4**

Write down the general solution of

$$X''(t) + X(t) = \cos t,$$

and choose the particular solution describing a particle that starts at time 0 with velocity  $v_0$  and displacement  $x_0$ .

**Exercise 5**

Obviously it is not *really* possible for vibrations to grow to arbitrarily large amplitude. The prediction of our mathematical model is that they will do so in the case  $p = 1$ . What features of the physical situation, omitted from our mathematical model, ought to be included if the model is to give a more realistic description of the case  $p = 1$ ?



In fact, the equation of motion used in this text is a good approximation only for small displacements; for large displacements some more complicated mathematical model would have to be used.

Nevertheless, the model we have been studying does correctly predict and describe the initial growth of the oscillations, before they become too violent.

### Postscript

There are some startling real-life examples of resonance—some useful and some disastrous! An example of the former type is an electrical circuit tuned to receive signals of one particular type—such as a radio transmission. The circuit (radio receiver) responds only to the stimulus of a particular frequency—the resonance frequency.

On the other hand, in 1940 the suspension bridge over the Tacoma gorge in North America was destroyed by winds which forced it to vibrate at its natural frequency. There is a famous film taken by a bystander which shows the oscillations building up until the bridge collapses. Some musical instruments illustrate resonance; if a suitable note is played near a violin, the corresponding string will vibrate in sympathy. Slightly different, but in the same context, the famous soprano Dame Nellie Melba was able to shatter a wine glass by singing an appropriate note.

## 11.7 Additional Exercises

### *Exercise 1*

Find the general solution of

$$f'' - 5f' + 4f = 0,$$

where  $f$  has domain  $R$ . (Cf. Exercise 11.2.2.)

### *Exercise 2*

What goes wrong when we apply the method of the preceding exercise to the equation

$$f'' + 2f' + f = 0,$$

where  $f$  has domain  $R$ ?



## Exercise 3

If the equation of motion is

$$X''(t) + X(t) = \sin \frac{1}{2}t \quad (t \in \mathbb{R}_0^+),$$

and the particle starts from rest at the origin, find the formula for the displacement  $X(t)$ .

## 11.8 Answers to Exercises

## Section 11.1

## Exercise 1

The velocity is the rate of change of position, and its value at time  $t$  is therefore given by the derivative of  $X$  at  $t$ ,

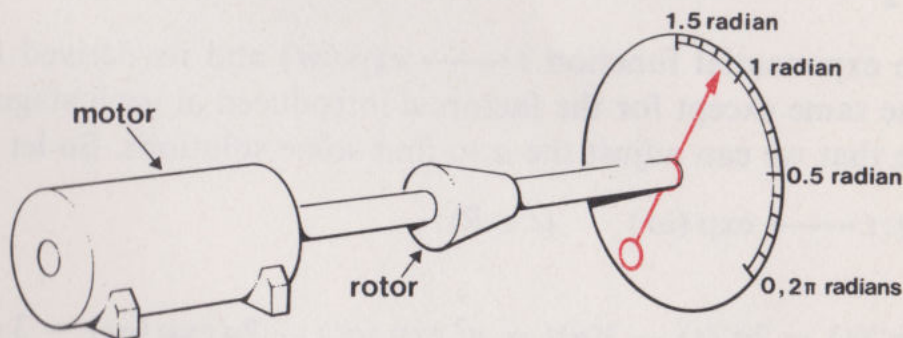
$$(\text{velocity at time } t) = X'(t).$$

The acceleration is the rate of change of velocity, and its value at time  $t$  is therefore given by the second derivative of  $X$  at  $t$ ,

$$(\text{acceleration at time } t) = X''(t).$$

## Exercise 2

Here, instead of a position on a line, it is the orientation of the rotor that we wish to describe. In place of the linear scale in the spring example, we have a scale measuring angles.



The motion can be described by a real function

$$f: t \longmapsto (\text{reading of pointer})$$

with the pointer reading in radians. This representation, however, has the disadvantage that  $f(t)$  not only jumps from  $2\pi$  to 0 as the pointer



passes (anti-clockwise) through zero, but also does not record the (net) magnitude of the rotation. To avoid these difficulties it is usual to work instead with the function

$$f: t \longmapsto (\text{reading of pointer} + 2\pi n),$$

where  $n$  is the net number of times the pointer has passed zero, counting anti-clockwise as positive. That is, every time the pointer passes zero in the anti-clockwise direction, we add a further  $2\pi$  to its subsequent readings, and every time it passes zero in the clockwise direction we subtract  $2\pi$ .

## Section 11.2

### Exercise 1

The equation can be written as

$$D^2g = g,$$

so that  $D^2$  has the same effect on  $g$  as multiplication by 1. This suggests the exponential function. If we try  $t \longmapsto \exp(at)$ , then we find that we require  $a^2 = 1$ , i.e.  $a = \pm 1$ . Hence two solutions are

$$t \longmapsto e^t \quad (t \in R) \quad \text{and} \quad t \longmapsto e^{-t} \quad (t \in R).$$

You should verify that both these functions are solutions of the differential equation  $g'' - g = 0$ .

### Exercise 2

Since the exponential function  $t \longmapsto \exp(at)$  and its derived functions are all the same except for the factors  $a$  introduced at each stage, there is still hope that we can adjust the  $a$  to find some solutions. So let us try

$$g: t \longmapsto \exp(at) \quad (t \in R);$$

then

$$\begin{aligned} g''(t) - 2g'(t) - 3(g)t &= a^2 \exp(at) - 2a \exp(at) - 3 \exp(at) \\ &= (a^2 - 2a - 3) \exp(at). \end{aligned}$$

This will be zero for all  $t$  if

$$a^2 - 2a - 3 = 0$$

i.e. if

$$(a - 3)(a + 1) = 0$$



i.e. if

$$a = 3 \text{ or } a = -1.$$

Thus two solutions are

$$g:t \longmapsto \exp(3t) \quad (t \in \mathbb{R});$$

$$g:t \longmapsto \exp(-t) \quad (t \in \mathbb{R}).$$

You should verify that both these functions are solutions of the differential equation.

### Section 11.3

#### Exercise 1

If

$$X(t) = \alpha \cos \omega t + \beta \sin \omega t,$$

then differentiation gives

$$X'(t) = -\alpha\omega \sin \omega t + \beta\omega \cos \omega t$$

and

$$\begin{aligned} X''(t) &= -\alpha\omega^2 \cos \omega t - \beta\omega^2 \sin \omega t \\ &= -\omega^2 X(t), \end{aligned}$$

so that

$$mX'' + sX = 0.$$

#### Exercise 2

To prove that the cosine and sine functions we are using are linearly independent, we have to show that

$$\alpha \cos + \beta \sin = 0 \text{ implies } \alpha = \beta = 0,$$

i.e. that

$$\alpha \cos \omega t + \beta \sin \omega t = 0 \quad \text{for all } t \in \mathbb{R}_0^+$$

implies

$$\alpha = \beta = 0.$$

One way to prove this is to consider a few special values of  $t$ . If we take  $t = 0$ , the equation to be satisfied reduces to  $\alpha \times 1 + \beta \times 0 = 0$ ; i.e.  $\alpha = 0$ : if we take  $t = \pi/2\omega$ , it reduces to  $\alpha \times 0 + \beta \times 1 = 0$ ; i.e.  $\beta = 0$ .

#### Exercise 3

(i) This operator is not linear: the  $f \times f$  term spoils it.



- (ii) This operator is linear.  
 (iii) This operator is not linear: the  $x \mapsto 2$  term spoils it.

#### Exercise 4

- (i) With  $f: x \mapsto 1$ , for example, we have

$$2H(f) = 2(D^4f - f \times f) = x \mapsto -2$$

but

$$H(2f) = D^4(2f) - (2f) \times (2f) = x \mapsto -4.$$

- (iii) With  $f: x \mapsto 1$  we have

$$2H(f) = 2(D^4f - (x \mapsto 2)) = x \mapsto -4$$

but

$$H(2f) = D^4(2f) - (x \mapsto 2) = x \mapsto -2.$$

#### Exercise 5

- (i) The equation is of the form  $Lf = 0$ , where  $L$  is a linear differential operator of degree 2 with constant coefficients. By the theorem, its kernel is therefore a vector space of dimension 2, and this kernel is the solution set of the equation.  
 (ii) It was shown in Exercise 11.2.1 that  $D^2f = f$  has the solutions

$$t \mapsto e^t \quad (t \in \mathbb{R})$$

and

$$t \mapsto e^{-t} \quad (t \in \mathbb{R}).$$

These solutions are linearly independent. (If  $\alpha e^t + \beta e^{-t} = 0$  for all  $t$ , then we have, putting  $t = 0$  and then  $t = 1$ ,  $\alpha + \beta = 0$  and  $\alpha e + \beta e^{-1} = 0$ , and these two equations taken together imply that  $\alpha = \beta = 0$ .)

- (iii) Taking the two linearly independent solutions given in (ii) as base vectors for the vector space, we obtain the general solution:

$$\alpha(t \mapsto e^t) + \beta(t \mapsto e^{-t}) \quad (t \in \mathbb{R}),$$

i.e.

$$t \mapsto \alpha e^t + \beta e^{-t} \quad (t \in \mathbb{R}),$$

where  $\alpha$  and  $\beta$  are arbitrary real numbers.



## Section 11.4

## Exercise 1

In this case

$$X(0) = x_0 \quad \text{and} \quad X'(0) = v_0,$$

so that

$$X:t \longmapsto x_0 \cos \omega t + \frac{v_0}{\omega} \sin \omega t \quad (t \in \mathbb{R}_0^+).$$

## Exercise 2

The number of oscillations per second is  $1/T$ , which is equal to  $\omega/2\pi$ . To increase this, alternatives (b) and (c) (increase  $s$  or decrease  $m$ ) are appropriate.

This result can also be understood on an intuitive basis directly from the mechanics: a light mass is moved rapidly by a strong spring, whereas a heavy mass attached to a weak spring will be sluggish.

## Section 11.5

## Exercise 1

Substituting the suggested form of  $X$  into the differential equation, we obtain the condition

$$m(-p^2k \cos pt) + sk \cos pt = F_0 \cos pt$$

i.e.

$$(-mp^2k + sk) \cos pt = F_0 \cos pt,$$

which is satisfied if

$$-mp^2k + sk = F_0.$$

Since  $p^2 \neq s/m$ , this equation can be solved for  $k$ , giving

$$k = \frac{F_0}{s - mp^2}.$$

The relevant particular solution of the equation is therefore

$$X:t \longmapsto \left( \frac{F_0}{s - mp^2} \right) \cos pt \quad (t \in \mathbb{R}_0^+).$$



*Exercise 2*

The given equation is like the previous one, but with a sine in place of a cosine. This suggests trying

$$X: t \longmapsto k \sin pt \quad (t \in R_0^+)$$

as a possible solution, and substitution in the differential equation shows that this function does satisfy the equation, where

$$k = \frac{1}{1 - p^2}.$$

A particular solution of the equation is thus

$$X: t \longmapsto \left( \frac{1}{1 - p^2} \right) \sin pt \quad (t \in R_0^+).$$

*Exercise 3*

The particular solution of  $L(X) = g$  is

$$t \longmapsto \left( \frac{F_0}{s - mp^2} \right) \cos pt \quad (t \in R_0^+).$$

(See Exercise 1.)

The complementary function, i.e. the kernel of  $L$ , is the general solution of  $mX'' + sX = 0$ , which we found in section 11.3 to be

$$t \longmapsto \alpha \cos \omega t + \beta \sin \omega t \quad (t \in R_0^+).$$

Putting these together, we find the general solution of  $L(X) = g$  to be

$$t \longmapsto \left( \frac{F_0}{s - mp^2} \right) \cos pt + \alpha \cos \omega t + \beta \sin \omega t \quad (t \in R_0^+),$$

provided  $p^2 \neq s/m$ .

*Exercise 4*

Differentiation of the supposed solution gives

$$X'(t) = - \left( \frac{pF_0}{s - mp^2} \right) \sin pt - \alpha \omega \sin \omega t + \beta \omega \cos \omega t,$$

$$X''(t) = - \left( \frac{p^2 F_0}{s - mp^2} \right) \cos pt - \alpha \omega^2 \cos \omega t - \beta \omega^2 \sin \omega t,$$



whence

$$mX''(t) + sX(t) = \left( \frac{-mp^2 F_0 \cos pt + sF_0 \cos pt}{s - mp^2} \right) = F_0 \cos pt.$$

as required.

### Exercise 5

We have already noted the analogy between the equation under consideration here, which has the form  $L(X) = g$ , and the matrix equation

$$A\underline{x} = \underline{b}.$$

A related problem is that of proving that if  $\underline{x}_0$  is a particular solution of  $A\underline{x} = \underline{b}$ , then the general solution of  $A\underline{x} = \underline{b}$  is

$$\underline{x} = \underline{x}_0 + (\text{general element of the kernel of } A).$$

The proof of this was given in section 7.4, and we repeat it here. Let  $\underline{x}_0$  be the particular solution of  $A\underline{x} = \underline{b}$ , and  $\underline{x}$  be any other solution, so that

$$A\underline{x} = \underline{b}$$

and

$$A\underline{x}_0 = \underline{b}.$$

Subtracting, and using the fact that  $A$  is a linear transformation, gives

$$A(\underline{x} - \underline{x}_0) = \underline{0},$$

so that  $\underline{x} - \underline{x}_0$  belongs to the kernel of  $A$ , i.e.  $\underline{x}$  is of the form

$$\underline{x}_0 + \text{an element of the kernel of } A.$$

Thus any solution of  $A\underline{x} = \underline{b}$  is of this form: in other words, the general solution has the form

$$\underline{x}_0 + (\text{general element of the kernel of } A).$$

To prove the corresponding result here we need little more than a change of notation. The given differential equation can be written as

$$L(X) = g$$

with

$$L: X \longmapsto mX'' + sX \quad \text{and} \quad g: t \longmapsto F_0 \cos pt.$$

We have seen already that one solution of this equation is the function

$$t \longmapsto \left( \frac{F_0}{s - mp^2} \right) \cos pt \quad (t \in R_0^+),$$



which we denote by  $X_0$ . Then if  $X$  is any solution of the differential equation, we have

$$LX = g$$

and

$$LX_0 = g.$$

Since  $L$  is a linear transformation, subtraction gives  $L(X - X_0) = O$ , so that  $X - X_0$  must belong to the kernel of  $L$ , that is, to the solution set of equation  $L(X) = O$ . We have seen in section 11.3 that this set comprises all functions of the form

$$t \longmapsto \alpha \cos \omega t + \beta \sin \omega t \quad (t \in R_0^+);$$

therefore  $X - X_0$  must be of this form, and so we have, by the definition of  $X_0$ ,

$$X:t \longmapsto \left( \frac{F_0}{s - mp^2} \right) \cos pt + \alpha \cos \omega t + \beta \sin \omega t \quad (t \in R_0^+),$$

for some  $\alpha$  and  $\beta$ . Thus every solution  $X$  of  $L(X) = g$  does have the form stated.

## Section 11.6

### Exercise 1

Starting as in the example, we have this time

$$\left( \frac{F_0}{s - mp^2} \right) + \alpha = x_0 \quad \text{and} \quad \beta\omega = v_0,$$

so that

$$\alpha = x_0 - \left( \frac{F_0}{s - mp^2} \right) \quad \text{and} \quad \beta = \frac{v_0}{\omega}.$$

The motion corresponds to the function

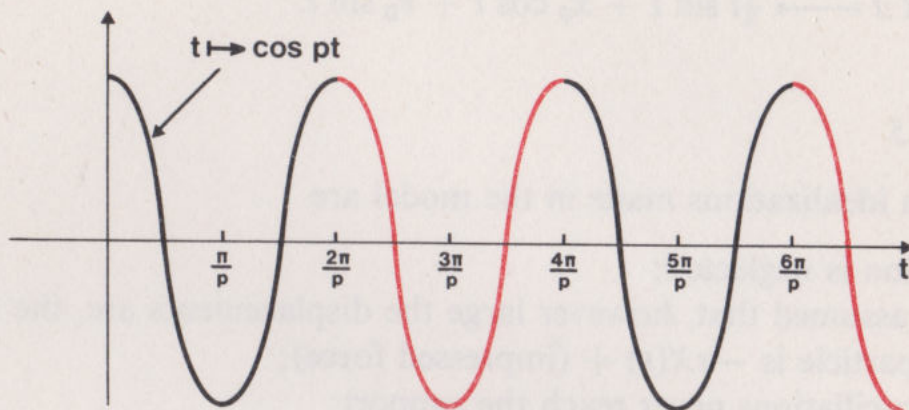
$$X:t \longmapsto \left( \frac{F_0}{s - mp^2} \right) \cos pt + \left( x_0 - \frac{F_0}{s - mp^2} \right) \cos \omega t + \frac{v_0}{\omega} \sin \omega t$$

$$(t \in R_0^+)$$



## Exercise 2

(i)



We see that the section of the graph in the interval  $\left[0, \frac{2\pi}{p}\right]$  repeats itself again and again and that no smaller interval has this property. The period is therefore  $2\pi/p$ .

- (ii) The  $\sin \frac{1}{2}t$  term has period  $4\pi$  and the  $\sin t$  term has period  $2\pi$ . So the complete function has period  $4\pi$ .

## Exercise 3

If

$$X(t) = \frac{1}{2}t \sin t,$$

then

$$X'(t) = \frac{1}{2} \sin t + \frac{1}{2}t \cos t,$$

$$X''(t) = \frac{1}{2} \cos t + \frac{1}{2} \cos t - \frac{1}{2}t \sin t,$$

and so

$$X''(t) + X(t) = \cos t \quad \text{as required.}$$

## Exercise 4

The general solution is

$$X:t \longmapsto \frac{1}{2}t \sin t + \alpha \cos t + \beta \sin t,$$

whence

$$X(0) = \alpha$$

$$X'(0) = \beta.$$



If the particle starts at time 0 with velocity  $v_0$  and displacement  $x_0$ , we have  $\alpha = x_0$ ,  $\beta = v_0$ , and

$$X:t \longmapsto \frac{1}{2}t \sin t + x_0 \cos t + v_0 \sin t.$$

### Exercise 5

The main idealizations made in the model are

- (i) friction is neglected;
- (ii) it is assumed that, however large the displacements are, the force on the particle is  $-sX(t)$  + (impressed force);
- (iii) the oscillations never reach the support.

Thus we are assuming that the particle does not hit any other objects (e.g. its support), that the properties of the spring do not change with time (in particular, it doesn't break) and that the force in the spring varies linearly with the displacement  $X(t)$ , no matter how violently it is stretched and compressed.

## Section 11.7

### Exercise 1

If we try

$$f:t \longmapsto \exp(ct) \quad (t \in R)$$

in the equation, the left-hand side becomes

$$\begin{aligned} f'' - 5f' + 4f &= t \longmapsto (c^2 e^{ct} - 5c e^{ct} + 4e^{ct}) & (t \in R) \\ &= (c^2 - 5c + 4)(t \longmapsto e^{ct}) & (t \in R). \end{aligned}$$

This equation is satisfied if the right-hand side reduces to the zero function, i.e. if

$$c^2 - 5c + 4 = 0$$

i.e.

$$c = 1 \text{ or } 4$$

Thus, the differential equation has the two solutions

$$t \longmapsto \exp t \quad \text{and} \quad t \longmapsto \exp 4t,$$

which can be verified to be linearly independent. Using them as base



vectors for the two-dimensional vector space of solutions, we obtain the general solution:

$$t \longmapsto \alpha \exp t + \beta \exp 4t \quad (t \in \mathbb{R}),$$

where  $\alpha$  and  $\beta$  are arbitrary real numbers.

### Exercise 2

Applying the same method as before, we find that the condition for  $t \longmapsto \exp(ct)$  to be a solution is

$$c^2 + 2c + 1 = 0.$$

This quadratic equation has only one solution,  $c = -1$ , and so the differential equation has only one solution of the form  $t \longmapsto \exp(ct)$ , namely

$$t \longmapsto \exp(-t) \quad (t \in \mathbb{R}).$$

We know, however, that the solution space is two-dimensional; therefore the single function displayed above is insufficient for a basis for the solution space. In this case, therefore, the method we have been using is inadequate. The difficulty can be overcome: you may like to have a try.

### Exercise 3

In Exercise 11.5.2 we found a particular solution of the equation

$$X''(t) + X(t) = \sin pt,$$

namely,

$$X: t \longmapsto \left( \frac{1}{1-p^2} \right) \sin pt \quad (t \in \mathbb{R}_0^+).$$

Here we have the same equation with  $p = \frac{1}{2}$ , and so a particular solution is

$$X: t \longmapsto \frac{4}{3} \sin \frac{1}{2}t \quad (t \in \mathbb{R}_0^+).$$

The general solution of the equation of motion, obtained by adding to this particular solution the general element of the kernel of the operator

$$X \longmapsto X'' + X,$$

is

$$X: t \longmapsto \frac{4}{3} \sin \frac{1}{2}t + \alpha \cos t + \beta \sin t \quad (t \in \mathbb{R}_0^+).$$



To use the initial conditions we note that

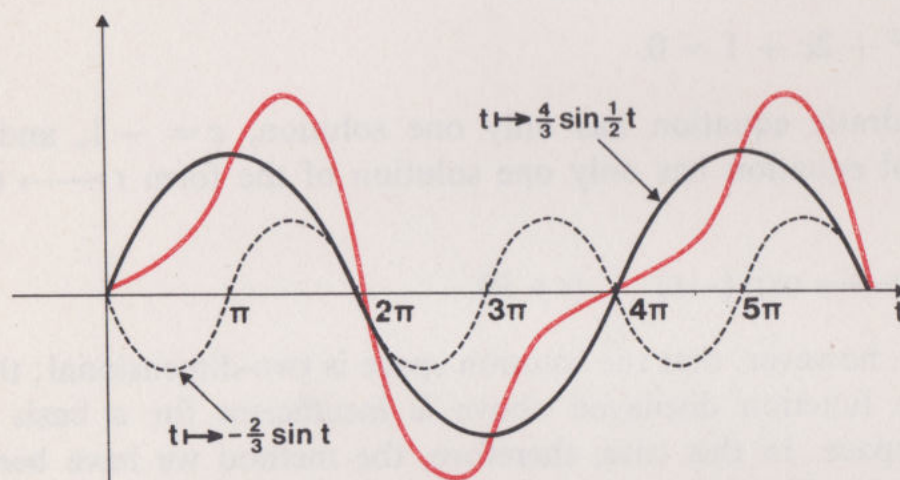
$$X(0) = \frac{4}{3} \sin 0 + \alpha \cos 0 + \beta \sin 0 = \alpha$$

and

$$X'(0) = \frac{2}{3} \cos 0 - \alpha \sin 0 + \beta \cos 0 = \frac{2}{3} + \beta,$$

so that  $\alpha = 0$  and  $\beta = -\frac{2}{3}$ , if the particle starts from rest at the origin. Substituting these values of  $\alpha$  and  $\beta$  in the above general solution, we obtained the required solution

$$X: t \longmapsto \frac{4}{3} \sin \frac{1}{2}t - \frac{2}{3} \sin t \quad (t \in \mathbb{R}_0^+).$$





# Index

A reference in **bold type** indicates that a definition of the term appears on that page.

- Addition of complex numbers **347**
- Addition of geometric vectors **114, 115**
- Addition of matrices **217**
- Amplitude **429**
- Anti-symmetric relation **61**
- Arbitrary constants **449**
- Argand diagram **346**
- Argument **343**
  - principal value of **343**
- Arrow **110**
- Associative operation **42**
- Augmented matrix **255**
- Axioms of a vector space **158**
- Axis, imaginary **346**
  - real **346**
  
- Back substitution **240**
- Base geometric vectors **128**
- Base vectors **128, 164**
- Basis **128, 164**
- Binary operation **37, 38**
  - induced **101**
- Binary relation **55**
- Bound, greatest lower **75**
  - least upper **76**
  - lower **75**
  - upper **74**
  
- Cartesian co-ordinates **334**
- Cartesian product **26**
- Closed operation **38**
- Codomain **9, 10**
- Column matrices **245**
- Column vectors **246**
- Commutative diagram **90**
- Commutative operation **40**
- Complement of a set **50**
- Complementary function **453**
- Complex exponential function **387**
- Complex function **373**
- Complex functions, composition of **399**
- Complex number **346**
  - conjugate of **355**
  - imaginary part of **346**
  - modulus of **355**
  - $n$ th root of **411**



# Index

- polar form of 350
- real part of 346
- square root of 407
- Complex numbers, addition of 347
  - division of 354
  - multiplication of 346
- Complex plane 346
- Composition of complex functions 389
- Composition of functions 18
- Conjugate of a complex number 355
- Constant coefficients 444
- Constant function 31
- Co-ordinates 124
  - cartesian 334
  - polar 334
- De Moivre's theorem 357
- Determinant 285
- Difference of functions 16
- Differential equation 155
  - order of 443
- Differential operator, linear 443
  - with constant coefficients 444
- Dimension of a vector space 165
- Dimension, infinite 165
- Dimensions 101, 103
- Dimension theorem 184
- Directed distance 431
- Direct methods 282
- Displacement 431
- Displacement-time graph 432
- Distributive operation 44
- Division of complex numbers 354
- Domain 9, 10
- Dot product 132
- Elementary matrix 269
- Elementary operation 238
- Elementary row operation 267
- Element of a set 2
- Empty set 4
- Equality of mappings 13
- Equality of matrices 205
- Equality of sets 4
- Equation, linear 201, 231
- Equation of motion 436
- Equations, simultaneous 231
- Equilibrium position 432
- Equivalence classes 67
- Equivalence relation 64
  - natural 69



# Index

- Equivalent systems 238
- Error vector 298
- Euler's formula 388
- Existence problem 248
- Explicit method 283
- Exponential function, complex 387
  
- Forced vibration 433, 451
- Free vibration 433, 451
- Function 10, 14, 27
  - complex exponential 387
  - constant 31
  - identity 221
  - inverse 22
  - many-one 14
  - one-one 14
  - square 348
  - zero 199
- Functions, composition of 18
  - difference of 16
  - product of 16
  - quotient of 16
  - sum of 15
  
- Gauss elimination method 238, 271
- General solution 439
- Geometric inversion 394
- Geometric vector 111
  - length of 119
  - modulus of 119
  - zero 117
- Geometric vectors, addition of 114, 115
  - linear combination of 124
  - scalar multiplication of 122
  - subtraction of 118
- Greatest lower bound 75
  
- Hasse diagrams 73
- Homomorphism 99
- Hooke's Law 434
- Hyper-plane 232
  
- Identity function 221
- Identity mapping 333
- Identity matrix 221
- Ill-conditioned equations 309
- Image 6, 10
- Imaginary axis 346
- Imaginary part of a complex number 346
- Indirect methods 295
- Induced binary operation 101



# Index

- Infinite dimension 165
- Inner product 132
- Intersection of sets 44, 46
- Invariant element 170
- Invariant set 170, 383
- Inverse function 22
- Inverse matrix 259, 274
- Inverse pair of points 394
- Isomorphism 99
- Iterative methods 295
  
- Joukowski aerofoil 402
  
- Kernel 183
  
- Least upper bound 76
- Left-distributive operation 44
- Length of a geometric vector 119
- Linear combination of geometric vectors 124
- Linear combination of vectors 162
- Linear differential operator 443
- Linear equations 201, 231
- Linearly dependent vectors 125, 126, 162
- Linearly independent vectors 125, 126, 162
- Linear transformation 175, 204
- Lower bound 75
  - greatest 75
  
- Many-many mapping 14
- Many-one function 14
- Mapping 5, 9, 14, 26
  - many-many 14
  - natural 69
  - one-many 14
  - reverse 20, 21
  - square root 408
- Mappings, equality of 13
- Mass 433
- Mathematical model 88
- Matrices, column 245
  - equality of 205
  - row 245
- Matrix 204
  - augmented 255
  - determinant of 285
  - elementary 269
  - identity 221
  - inverse of 259, 274
  - non-singular 259
  - order of 259
  - rank of 252, 279



# Index

- scalar multiplication of **216**
- singular **259**
- sparse **295**
- square **221**
- zero **222**
- Matrix addition **217**
- Matrix multiplication **214, 218**
- Matrix subtraction **217**
- Modulus of a complex number **355**
- Modulus of a geometric vector **119**
- Morphism **96**
- Morphism of a vector space **175**
- Multiplication of complex numbers **346**
- Multiplication of matrices **214, 218**
  
- $N$ -ary operation **48**
- $n$ -ary relation **55**
- Natural equivalence relation **69**
- Natural mapping **69**
- Newton's second law **433**
- Non-singular matrix **259**
- Norm **304**
- $n$ th root of a complex number **411**
- Null space **183**
  
- One-many mapping **14**
- One-one function **14**
- Operation, associative **42**
  - binary **37, 38**
  - closed **38**
  - commutative **40**
  - distributive **44**
  - elementary **238, 267**
  - $N$ -ary **48**
  - ternary **48**
  - unary **48**
- Operator, linear differential **443**
- Order of a differential equation **443**
- Order of matrix **259**
- Order relation **64**
  - partial **72**
  - total **72**
- Ordered lists **152**
- Ordered pair **25**
- Ordered set **72**
  
- Partial ordering relation **72**
- Particle **431**
- Particular solution **247**
- Particular solution of a differential equation **453, 454**



# Index

- Partition of a set 65
- Periodic motion 448
- Period of motion 449
- Polar co-ordinates 334
- Polar form of a complex number 350
- Principal value of the argument 343
- Product of functions 16
- Projection 132
- Proper subset 4
  
- Quadrant 346
- Quotient of functions 16
- Quotient set 68
  
- Rank of a matrix 252, 279
- Real axis 346
- Real part of a complex number 346
- Reflexive relation 58
- Relation 50, 53, 55
  - anti-symmetric 61
  - binary 55
  - equivalence 64
  - $n$ -ary 55
  - order 64
  - reflexive 58
  - solution set of 54
  - symmetric 60
  - ternary 55
  - transitive 62
- Resonance 430, 457
- Reverse mapping 20, 21
- Right distributive operation 44
- Rotation matrices 218
- Row matrices 245
- Row vectors 246
  
- Scalar 122
- Scalar multiplication of geometric vectors 122
- Scalar multiplication of a matrix 216
- Scalar product 132
- Set 1
  - complement of 50
  - element of 2
  - empty 4
  - ordered 72
  - partition of 65
  - quotient 69
  - solution 232, 233
- Sets, intersection of 44, 46
  - union of 44, 45
- Simpsons rule 157



# Index

- Simultaneous equations 231
- Singular matrix 259
- Solution, particular 247
  - unique 237
- Solution set 232, 233
- Solution set of a relation 54
- Span 124, 164
- Sparse matrix 295
- Square function 348
- Square matrix 221
- Square root mapping 408
- Square root of a complex number 407
- Stiffness 434
- Sub-matrices 245
- Subset 4
  - proper 4
- Subtraction of geometric vectors 118
- Subtraction of matrices 217
- Sum of functions 15
- Symmetric relation 60
  
- Ternary operation 48
- Ternary relation 55
- Total ordering relation 72
- Transitive relation 62
- Translation 113
- Triangle inequality 120
  
- Unary operation 48
- Union of sets 44, 45
- Uniqueness problem 248
- Unique solution 237
- Unit of measurement 101
- Upper bound 74
  - least 76
  
- Variable 12
- Vector 158
  - error 298
  - geometric 111
  - zero 159, 160
- Vectors, column 246
  - linear combination of 162
  - linearly dependent 162
  - linearly independent 162
  - row 246
- Vector space 158
  - axioms of 158
  - dimension of 165
  - morphism of 175



# Index

Vector subspace 178  
Vibration, forced 433, 451  
    free 433, 451  
  
w-plane 377  
  
Zero function 199  
Zero geometric vector 117  
Zero matrix 222  
Zero vector 159, 160  
z-plane 377



2

2



# AN INTRODUCTION TO CALCULUS AND ALGEBRA

## *Volume 3 Algebra*

This third volume develops concepts of algebra and begins by repeating some of the early material of Volume 1 common to both algebra and calculus. Starting with *sets* and *mappings*, it then introduces *binary operations*, *relations* and *morphisms*. Five chapters are devoted to linear algebra, in particular to *geometric vectors*, *vector spaces* and *matrices*, and two chapters to *complex numbers* and *elementary complex functions*. The last of its eleven chapters uses a discussion of *second order differential equations* to link together concepts in both calculus and algebra which have appeared in the three volumes.

This is the third of three volumes presenting some of the essential concepts of mathematics, a few important proofs (usually in outline), together with exercises designed to reinforce the understanding of the concepts and develop the beginnings of technical skill.

A particular feature of these volumes is the use of two-colour printing to emphasize important concepts and definitions and to heighten the impact of diagrams. Exercises are provided throughout, and detailed answers (with additional comment) are given at the end of each chapter.

The selection of material has been made with the needs of students of other subjects particularly in mind. The material presented is, therefore, suited to a wide class of reader and will provide both a modern introduction and a handy reference to two important areas of mathematics: Calculus and Algebra.

Volume 1 Background to Calculus

Volume 2 Calculus Applied

Volume 3 Algebra



THE OPEN UNIVERSITY PRESS

SBN 335 00004 5